SUBSTITUTE SPECIFICATION

APPARATUS AND METHOD FOR AUTOMATED PROTEIN DESIGN

This application claims the benefit of U.S.S.N.s 60/043,464, filed April 11, 1997, 60/054,678, filed August 4, 1997, 60/061,097, filed October 3, 1997, 06/087,561, filed June 1, 1998, and is a continuing application of U.S.S.N. 09/058,459, filed April 10, 1998.

FIELD OF THE INVENTION

The present invention relates to an apparatus and method for quantitative protein design and optimization.

BACKGROUND OF THE INVENTION

De novo protein design has received considerable attention recently, and significant advances have been made toward the goal of producing stable, well-folded proteins with novel sequences. Efforts to design proteins rely on knowledge of the physical properties that determine protein structure, such as the patterns of hydrophobic and hydrophilic residues in the sequence, salt bridges and hydrogen bonds, and secondary structural preferences of amino acids. Various approaches to apply these principles have been attempted. For example, the construction of α-helical and β-sheet proteins with native-like sequences was attempted by individually selecting the residue required at every position in the target fold (Hecht, *et al.*, Science **249**:884-891 (1990); Quinn, *et al.*, Proc. Natl. Acad. Sci USA **91**:8747-8751 (1994)). Alternatively, a minimalist approach was used to design helical proteins, where the simplest possible sequence believed to be consistent with the folded structure was generated (Regan, *et al.*, Science **241**:976-978 (1988); DeGrado, *et al.*, Science **243**:622-628 (1989); Handel, *et al.*, Science **261**:879-885 (1993)), with varying degrees of success. An experimental method that relies on the hydrophobic and polar (HP) pattern of a sequence was developed where a library of sequences with the correct pattern for a four helix bundle was generated by random mutagenesis (Kamtekar, *et al.*, Science **262**:1680-1685 (1993)). Among non de novo approaches, domains of naturally occurring proteins have been modified or coupled together to achieve a desired tertiary organization (Pessi, *et al.*, Nature **362**:367-369 (1993); Pomerantz, *et al.*, Science **267**:93-96 (1995)).

Though the correct secondary structure and overall tertiary organization seem to have been attained by several of the above techniques, many designed proteins appear to lack the structural specificity of native proteins. The complementary geometric arrangement of amino acids in the folded protein is the root of this specificity and is encoded in the sequence.

Several groups have applied and experimentally tested systematic, quantitative methods to protein design with the goal of developing general design algorithms (Hellinga, *et al.*, J. Mol. Biol. **222**: 763-785 (1991); Hurley, *et al.*, J. Mol. Biol. **224**:1143-1154 (1992); Desjarlaisl, *et al.*, Protein Science **4**:2006-2018 (1995); Harbury, *et al.*, Proc. Natl. Acad. Sci. USA **92**:8408-8412 (1995); Klemba, *et al.*, Nat. Struc. Biol. **2**:368-373 (1995); Nautiyal, *et al.*, Biochemistry **34**:11645-11651 (1995); Betzo, *et al.*, Biochemistry **35**:6955-6962 (1996); Dahiyat, *et al.*, Protein Science **5**:895-903 (1996); Jones, Protein Science **3**:567-574 (1994); Konoi, *et al.*, Proteins: Structure, Function and Genetics **19**:244-255 (1994)). These algorithms consider the spatial positioning and steric complementarity of side chains by explicitly modeling the atoms of sequences under consideration. To date, such techniques have typically focused on designing the cores of proteins and have scored sequences with van der Waals and sometimes hydrophobic solvation potentials.

Recent studies using coiled coils have demonstrated that core side-chain packing can be combined with explicit backbone flexibility (Harbury et al., PNAS USA 92:8408-8412 (1995); Offer & Sessions, J. Mol. Biol. 249:967-987 (1995). In these cases, the goal was to search for backbone coordinates that satisfied a fixed amino acid sequence.

In addition, the qualitative nature of many design approaches has hampered the development of improved, second generation, proteins because there are no objective methods for learning from past design successes and failures.

Thus, it is an object of the invention to provide computational protein design and optimization via an objective, quantitative design technique implemented in connection with a general purpose computer.

SUMMARY OF THE INVENTION

In accordance with the objects outlined above, the present invention provides methods executed by a computer under the control of a program, the computer including a memory for storing the program. The method comprising the steps of receiving a protein backbone structure with variable residue positions, establishing a group of potential rotamers for each of the variable residue positions, wherein at least one variable residue position has rotamers from at least two different amino acid side chains, and analyzing the interaction of each of the rotamers with all or part of the remainder of the protein backbone structure to generate a set of optimized protein sequences. The methods further comprise classifying each variable residue position as either a core, surface or boundary residue. The analyzing step may include a Dead-End Elimination (DEE) computation. Generally, the analyzing step includes the use of at least one scoring function selected from the group consisting of a Van der Waals potential scoring function, a hydrogen bond potential scoring function, an atomic solvation scoring function, a secondary structure propensity scoring function and an electrostatic scoring function. The methods further comprise altering the protein backbone prior to the analysis, comprising altering at least one supersecondary structure parameter value. The methods may further comprise generating a rank ordered list of additional optimal sequences from the globally optimal

protein sequence. Some or all of the protein sequences from the ordered list may be tested to produce potential energy test results.

In an additional aspect, the invention provides nucleic acid sequences encoding a protein sequence generated by the present methods, and expression vectors and host cells containing the nucleic acids.

In a further aspect, the invention provides a computer readable memory to direct a computer to function in a specified manner, comprising a side chain module to correlate a group of potential rotamers for residue positions of a protein backbone model, and a ranking module to analyze the interaction of each of said rotamers with all or part of the remainder of said protein to generate a set of optimized protein sequences. The memory may further comprise an assessment module to assess the correspondence between potential energy test results and theoretical potential energy data.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a general purpose computer configured in accordance with an embodiment of the invention.

Figure 2 illustrates processing steps associated with an embodiment of the invention.

Figure 3 illustrates processing steps associated with a ranking module used in accordance with an embodiment of the invention. After any DEE step, any one of the previous DEE steps may be repeated. In addition, any one of the DEE steps may be eliminated; for example, original singles DEE (step 74) need not be run.

Figure 4 depicts the protein design automation cycle.

Figure 5 depicts the helical wheel diagram of a coiled coil. One heptad repeat is shown viewed down the major axes of the helices. The **a** and **d** positions define the solvent-inaccessible core of the molecule (Cohen & Parry, 1990, Proteins, Structure, Function and Genetics 7:1-15).

Figures 6A and 6B depict the comparison of simulation cost functions to experimental Tm's. Figure 6A depicts the initial cost function, which contains only a van der Waals term for the eight PDA peptides. Figure 6B depicts the improved cost function containing polar and nonpolar surface area terms weighted by atomic solvation parameters derived from QSAR analysis; 16 cal/mol/Å^2 favors hydrophobic surface burial.

Figure 7 shows the rank correlation of energy predicted by the simulation module versus the combined activity score of λ repressor mutants (Lim, *et al.*, J. Mol. Biol. **219**:359-376 (1991); Hellinga, *et al.*, Proc. Natl. Acad. Sci. USA **91**:5803-5807 (1994)).

Figure 8 shows the sequence of pda8d (SEQ ID NO:2) aligned with the second zinc finger of Zif268 (SEQ D NO:1). The boxed ositions were designed using the sequence selection algorithm. The coordinates of PDB record 1zaa (Paveletch, *et al.*, Science **252**:809-817 (1991)) from residues 33-60 were used as the structure template. In our numbering, position 1 corresponds to 1zaa position 33.

Figures 9A and 9B shows the NMR spectra and solution secondary structure of pda8d from Example 3. Figure 9A is the TOCSY Hα-HN fingerprint region of pda8d. Figure 9B is the NMR NOE connectivities of pda8d. Bars represent unambiguous connectivities and the bar thickness of the sequential connections is indexed to the intensity of the resonance.

Figures 10A and 10B depict the secondary structure content and thermal stability of α90, α85, α70 and α107. Figure 10A depicts the far UV spectra (circular dichroism). Figure 10B depicts the thermal denaturation monitored by CD.

Figure 11 depicts the sequence of FSD-1 (SEQ ID NO:3) of Example 5 aligned with the second zinc finger of Zif268 (SEQ ID NO:1). The bar at the top of the figure shows the residue position classifications: solid bars indicate core positions, hatched bars indicate boundary positions and open bars indicate surface positions. The alignment matches positions of FSD-1 (SEQ ID NO:3) to the corresponding backbone template positions of Zif268 (SEQ ID NO:1). Of the six identical positions (21%) between FSD-1 (SEQ ID NO:3) and Zif268 (SEQ ID NO:1), four are buried (Ile7, Phe12, Leu18 and Ile22). The zinc binding residues of Zif268 are boxed. Representative non-optimal sequence solutions (SEQ ID NOS:4 –22) determined using a Monte Carlo simulated annealing protocol are shown with their rank. Vertical lines indicate identity with FSD-1 (SEQ ID NO:3). The symbols at the bottom the figure show the degee of sequence conservation for each residue position computed across the top 1000 sequences: filled circles indicate greater than 99% conservation, half-filled circles indicate conservation between 90 and 99%, open circles indicate conservation between 50 and 90%, and the absence of symbol indicates less than 50% conservation. The consensus sequence determined by choosing the amino acid with the highest occurrence at each position is identical to the sequence of FSD-1 (SEQ ID NO:3).

Figure 12 is a schematic representation of the minimum and maximum quantities (defined in Eq. 24 to 27) that are used to construct speed enhancements. The minima and maxima are utilized directly to find the $\overline{i_u j_v}_{mb}$ pair and for the comparison of extrema. The differences between the quantities, denoted with arrows, are used to construct the q_{rs} and q_{uv} metrics.

Figures 13A, 13B, 13C, 13D, 13E and 13F depicts the areas involved in calculating the buried and exposed areas of Equations 18 and 19. The dashed box is the protein template, the heavy solid lines correspond to three rotamers at three different residue positions, and the lighter solid lines correspond to surface areas. a) $A^0_{i,t3}$ for each rotamer. b) $A_{i,t}$ for each rotamer. c) $(A^0_{i,t3} - A_{i,t})$ summed over the three residues. The upper residue does not bury any area against the template except that buried in the tri-peptide state $A^0_{i,t3}$. d) $A_{i,j,t}$ for one pair of rotamers. e) The area buried between

rotamers, $(A_{i,t} + A_{j,t} - A_{i,j,t})$, for the same pair of rotamers as in (d). f) The area buried between rotamers, $(A_{i,t} + A_{j,t} - A_{i,j,t})$, summed over the three pairs of rotamers. The area b intersected by all three rotamers is counted twice and is indicated by the double lines. The buried area calculated by Equation 18 is the area buried by the template, represented in (c), plus *s* times the area buried between rotamers, represented in (f). The scaling factor *s* accounts for the over-counting shown by the double lines in (f). The exposed area calculated by Equation 19 is the exposed are in the presence of the template, represented in (b), minus *s* times the area buried between rotamers, represented in (f).

Figures 14A, 14B, 14C and 14D depict several super-secondary structure parameters for α/β proteins. The definitions are similar to those previously developed for α/β proteins (Janin & Chothia, J Mol Biol 143:95–128 (1980); Cohen et al., J Mol Biol 156:821–862 (1982)). The helix center is defined as the average C_α position of the residues in the helix. The helix axis is defined as the principal moment of the C_α atoms of the residues in the helix. (Chothia et al., Proc Natl Acad Sci USA 78:4146–4150 (1981); J Mol Biol 145:215–250 (1981). The strand axis is defined as the average of the least-squares lines fit through the midpoints of sequential C_α positions of two central β-strands. The sheet plane is defined as the least-squares plane fit through the C_α positions of the residues of the sheet. The sheet axis is defined as the vector perpendicular to the sheet plane that passes through the helix center. Ω is the angle between the strand axis and the helix axis after projection onto the sheet plane; θ is the angle between the helix axis and the sheet plane; h is the distance between the helix center and the sheet plane; σ is the rotation angle about the helix axis. The super-secondary structure parameter values for native Gβ1 are Ω = -26.49°, θ = 3.20°, h = 10.04 Å and σ = 0°.

Figure 15 depicts the Far-UV CD spectra of Gβ1 and the most perturbed of the Δh-series mutants, $\Delta h_{0.9}$[+1.50Å]. $\Delta h_{0.9}$[-1.50Å] and $\Delta h_{1.0}$[+1.50Å] have CD spectra similar to $\Delta h_{0.9}$[+1.50Å], while the remaining mutants have CD spectra similar to Gβ1.

Figure 16 depicts the thermal denaturation of Gβ1 and the Δh-series mutants monitored by CD at 218 nm.

Figures 17A, 17B, 17C and 17D depict four supersecondary structure parameters for β/β protein interactions. Figures 17A and 17B are relevant to β barrel proteins; Figure 17C is relevant to β-sheet interactions. Figure 17A shows only three strands, and depicts R, the barrel radius; α, the tilt of the strands relative to the barrel axis; a, the distance from C^α to C^α along the strands; and b, the interstrand distance. Figure 17B shows the twist and coiling angles of the β-sheet, with residues A, B and C from one strand, residues D, E and F in strand 2, and residues G, H and I from strand 3. The circles represent the positions of the residues when projected onto the surface of the barrel. In this case, θ is the mean twist of the β-sheet about an axis perpendicular to the strand direction. τ is the mean twist of the β-sheet about an axis parallel to the strand direction. ε is the mean coiling of the β-sheet along the strands. η is the mean coiling of the β-sheet along a line perpendicular to the strands. Figure 17C depicts two β-sheets, with the chain direction being shown with arrows. Figure 17D

depicts two β-sheets of distance h with angle θ between the average strand vectors. There is also φ, perpendicular to vectors defining θ.

Figures 18A, 18B, 18C and 18D depict four supersecondary structure parameters α/α supersecondary structure parameters for α/α interactions. d is the distance between the helices and θ is the angle between the axes of the helices. σ is defined as the rotation around the helix axis. Ω is the angle between two strand axes after projection onto a plane. In Figures 18C and 18D, the dark circle represents a view of the helix from the end.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to the quantitative design and optimization of amino acid sequences, using an "inverse protein folding" approach, which seeks the optimal sequence for a desired structure. Inverse folding is similar to protein design, which seeks to find a sequence or set of sequences that will fold into a desired structure. These approaches can be contrasted with a "protein folding" approach which attempts to predict a structure taken by a given sequence.

The general preferred approach of the present invention is as follows, although alternate embodiments are discussed below. A known protein structure is used as the starting point. The residues to be optimized are then identified, which may be the entire sequence or subset(s) thereof. The side chains of any positions to be varied are then removed. The resulting structure consisting of the protein backbone and the remaining sidechains is called the template. Each variable residue position is then preferably classified as a core residue, a surface residue, or a boundary residue; each classification defines a subset of possible amino acid residues for the position (for example, core residues generally will be selected from the set of hydrophobic residues, surface residues generally will be selected from the hydrophilic residues, and boundary residues may be either). Each amino acid can be represented by a discrete set of all allowed conformers of each side chain, called rotamers. Thus, to arrive at an optimal sequence for a backbone, all possible sequences of rotamers must be screened, where each backbone position can be occupied either by each amino acid in all its possible rotameric states, or a subset of amino acids, and thus a subset of rotamers.

Two sets of interactions are then calculated for each rotamer at every position: the interaction of the rotamer side chain with all or part of the backbone (the "singles" energy, also called the rotamer/template or rotamer/backbone energy), and the interaction of the rotamer side chain with all other possible rotamers at every other position or a subset of the other positions (the "doubles" energy, also called the rotamer/rotamer energy). The energy of each of these interactions is calculated through the use of a variety of scoring functions, which include the energy of van der Waal's forces, the energy of hydrogen bonding, the energy of secondary structure propensity, the energy of surface area solvation and the electrostatics. Thus, the total energy of each rotamer interaction, both with the backbone and other rotamers, is calculated, and stored in a matrix form.

The discrete nature of rotamer sets allows a simple calculation of the number of rotamer sequences to be tested. A backbone of length n with m possible rotamers per position will have m^n possible rotamer sequences, a number which grows exponentially with sequence length and renders the calculations either unwieldy or impossible in real time. Accordingly, to solve this combinatorial search problem, a "Dead End Elimination" (DEE) calculation is performed. The DEE calculation is based on the fact that if the worst total interaction of a first rotamer is still better than the best total interaction of a second rotamer, then the second rotamer cannot be part of the global optimum solution. Since the energies of all rotamers have already been calculated, the DEE approach only requires sums over the sequence length to test and eliminate rotamers, which speeds up the calculations considerably. DEE can be rerun comparing pairs of rotamers, or combinations of rotamers, which will eventually result in the determination of a single sequence which represents the global optimum energy.

Once the global solution has been found, a Monte Carlo search may be done to generate a rank-ordered list of sequences in the neighborhood of the DEE solution. Starting at the DEE solution, random positions are changed to other rotamers, and the new sequence energy is calculated. If the new sequence meets the criteria for acceptance, it is used as a starting point for another jump. After a predetermined number of jumps, a rank-ordered list of sequences is generated.

The results may then be experimentally verified by physically generating one or more of the protein sequences followed by experimental testing. The information obtained from the testing can then be fed back into the analysis, to modify the procedure if necessary.

Thus, the present invention provides a computer-assisted method of designing a protein. The method comprises providing a protein backbone structure with variable residue positions, and then establishing a group of potential rotamers for each of the residue positions. As used herein, the backbone, or template, includes the backbone atoms and any fixed side chains. The interactions between the protein backbone and the potential rotamers, and between pairs of the potential rotamers, are then processed to generate a set of optimized protein sequences, preferably a single global optimum, which then may be used to generate other related sequences.

Figure 1 illustrates an automated protein design apparatus 20 in accordance with an embodiment of the invention. The apparatus 20 includes a central processing unit 22 which communicates with a memory 24 and a set of input/output devices (e.g., keyboard, mouse, monitor, printer, etc.) 26 through a bus 28. The general interaction between a central processing unit 22, a memory 24, input/output devices 26, and a bus 28 is known in the art. The present invention is directed toward the automated protein design program 30 stored in the memory 24.

The automated protein design program 30 may be implemented with a side chain module 32. As discussed in detail below, the side chain module establishes a group of potential rotamers for a selected protein backbone structure. The protein design program 30 may also be implemented with a ranking module 34. As discussed in detail below, the ranking module 34 analyzes the interaction of rotamers with the protein backbone structure to generate optimized protein sequences. The protein

design program 30 may also include a search module 36 to execute a search, for example a Monte Carlo search as described below, in relation to the optimized protein sequences. Finally, an assessment module 38 may also be used to assess physical parameters associated with the derived proteins, as discussed further below.

The memory 24 also stores a protein backbone structure 40, which is downloaded by a user through the input/output devices 26. The memory 24 also stores information on potential rotamers derived by the side chain module 32. In addition, the memory 24 stores protein sequences 44 generated by the ranking module 34. The protein sequences 44 may be passed as output to the input/output devices 26.

The operation of the automated protein design apparatus 20 is more fully appreciated with reference to Fig. 2. Fig. 2 illustrates processing steps executed in accordance with the method of the invention. As described below, many of the processing steps are executed by the protein design program 30. The first processing step illustrated in Fig. 2 is to provide a protein backbone structure (step 50). As previously indicated, the protein backbone structure is downloaded through the input/output devices 26 using standard techniques.

The protein backbone structure corresponds to a selected protein. By "protein" herein is meant at least two amino acids linked together by a peptide bond. As used herein, protein includes proteins, oligopeptides and peptides. The peptidyl group may comprise naturally occurring amino acids and peptide bonds, or synthetic peptidomimetic structures, i.e. "analogs", such as peptoids (see Simon *et al.*, PNAS USA **89**(20):9367 (1992)).. The amino acids may either be naturally occuring or non-naturally occuring; as will be appreciated by those in the art, any structure for which a set of rotamers is known or can be generated can be used as an amino acid. The side chains may be in either the (R) or the (S) configuration. In a preferred embodiment, the amino acids are in the (S) or L-configuration.

The chosen protein may be any protein for which a three dimensional structure is known or can be generated; that is, for which there are three dimensional coordinates for each atom of the protein. Generally this can be determined using X-ray crystallographic techniques, NMR techniques, de novo modelling, homology modelling, etc. In general, if X-ray structures are used, structures at 2Å resolution or better are preferred, but not required.

The proteins may be from any organism, including prokaryotes and eukaryotes, with enzymes from bacteria, fungi, extremeophiles such as the archebacteria, insects, fish, animals (particularly mammals and particularly human) and birds all possible.

Suitable proteins include, but are not limited to, industrial and pharmaceutical proteins, including ligands, cell surface receptors, antigens, antibodies, cytokines, hormones, and enzymes. Suitable classes of enzymes include, but are not limited to, hydrolases such as proteases, carbohydrases, lipases; isomerases such as racemases, epimerases, tautomerases, or mutases; transferases,

kinases, oxidoreductases, and phophatases. Suitable enzymes are listed in the Swiss-Prot enzyme database.

Suitable protein backbones include, but are not limited to, all of those found in the protein data base compiled and serviced by the Brookhaven National Lab.

Specifically included within "protein" are fragments and domains of known proteins, including functional domains such as enzymatic domains, binding domains, etc., and smaller fragments, such as turns, loops, etc. That is, portions of proteins may be used as well. In addition, protein variants, i.e. non-naturally occuring variants, may be used.

Once the protein is chosen, the protein backbone structure is input into the computer. By "protein backbone structure" or grammatical equivalents herein is meant the three dimensional coordinates that define the three dimensional structure of a particular protein. The structures which comprise a protein backbone structure (of a naturally occuring protein) are the nitrogen, the carbonyl carbon, the α-carbon, and the carbonyl oxygen, along with the direction of the vector from the α-carbon to the β-carbon.

The protein backbone structure which is input into the computer can either include the coordinates for both the backbone and the amino acid side chains, or just the backbone, i.e. with the coordinates for the amino acid side chains removed. If the former is done, the side chain atoms of each amino acid of the protein structure may be "stripped" or removed from the structure of a protein, as is known in the art, leaving only the coordinates for the "backbone" atoms (the nitrogen, carbonyl carbon and oxygen, and the α-carbon, and the hydrogens attached to the nitrogen and α-carbon).

In a preferred embodiment, the protein backbone structure is altered prior to the analysis outlined below. In this embodiment, the representation of the starting protein backbone structure is reduced to a description of the spatial arrangement of its secondary structural elements. The relative positions of the secondary structural elements are defined by a set of parameters called supersecondary structure parameters. These parameters are assigned values that can be systematically or randomly varied to alter the arrangement of the secondary structure elements to introduce explicit backbone flexibility. The atomic coordinates of the backbone are then changed to reflect the altered supersecondary structural parameters, and these new coordinates are input into the system for use in the subsequent protein design automation.

Basically, a protein is first parsed into a collection of secondary structural elements which are then abstracted into geometrical objects. For example, as more fully outlined below, an α-helix is represented by its helical axis and geometric center. The relative orientation and distance between these objects are summarized as super-secondary structure parameters. Concerted backbone motion can be introduced by simply modulating a protein's super-secondary structure parameter values. Accordingly, when all or part of the backbone is to be altered, the portion to be altered is classified as belonging to a particular supersecondary structure element, i.e. α/β, α/α or β/β, and then

the supersecondary structural elements as outlined below are altered. As will be appreciated by those in the art, these elements need not be covalently linked, i.e. part of the same protein; for example, this can be done to evaluate protein-protein interactions.

As will be appreciated by those in the art, it is possible to alter the backbone of certain positions, while retaining either a particular amino acid (which can be "floated", as outlined below) or a particular rotamer at the position; alternatively, both the backbone can be moved and the amino acid side chain can be optimized as outlined herein. Similarly, the backbone can be held constant and only the amino acid side chains are optimized. Combinations of any of these at any position may be done. In general, when supersecondary structural parameters are altered, this is done on more than one amino acid, i.e. the backbone atoms of a plurality of amino acids that contribute to the secondary structure are moved.

As will be appreciated by those in the art, there are a wide variety of different supersecondary structure parameters that can be used. Super-secondary structure parameterization has been described for fold classes that include α/α (Crick FHC The Fourier transform of a coiled-coil. *Acta Crystallogr 6*:685–689 (1953a); Crick FHC. The packing of α-helices. *Acta Crystallogr 6*:689–697 (1953b); Chothia et al., *Proc Natl Acad Sci USA 78*:4146–4150 (1981) "Relative orientation of close-packed β-pleated sheets in proteins" and Chothia et al., *J Mol Biol 145*:215–250 (1981) "Helix to helix packing in proteins"; Chou, et al. Energetics of the structure of the four-α-helix bundle in proteins. *Proc Natl Acad Sci USA 85*:4295–4299 (1988); Murzin AG, Finkelstein AV. General architecture of the α-helical globule. *J Mol Biol 204*:749–769 (1988); Presnell SR, Cohen FE. Topological distribution of four-α-helix bundles. *Proc Natl Acad Sci USA 86*:6592–6596 (1989); Harris et al. Four helix bundle diversity in globular proteins. *J Mol Biol 236*:1356–1368 (1994), α/β (Chothia et al., Structure of proteins: packing of α-helices and pleated sheets. *Proc Natl Acad Sci USA 74*:4130–4134 (1977); Janin & Chothia, 1980 Packing of α-helices onto β-pleated sheets and the anatomy of *α/β* proteins. *J Mol Biol 143*:95–128; Cohen et al., 1982, Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J Mol Biol 156*:821–862; Chou et al., 1985, Interactions between an α-helix and β-sheet energetics of *α/β* packing in proteins. *J Mol Biol 186*:591–609, and β/β (Cohen et al., Analysis and prediction of protein β-sheet structures by a combinatorial approach. *Nature 285*:378–382 (1980); Cohen et al., Analysis of the tertiary structure of protein β-sheet sandwiches. *J Mol Biol 148*:253–272 (1981); Chothia & Janin, Relative orientation of close-packed β-pleated sheets in proteins. *Proc Natl Acad Sci USA 78*:4146–4150 (1981); Chothia & Janin, *Proc Natl Acad Sci USA 78*:3955–3965 (1982) Orthogonal packing of β-pleated sheets in proteins; Chou et al., *J Mol Biol 188*:641–649 (1986) "Interactions between two β-sheets energetics of *β/β* packing in proteins"; Laster et al., *Proc Natl Acad Sci USA 85*:3338–3342 (1988)"Structure principles of parallel β-barrels in proteins"; Murzin et al., *J Mol Biol 236*:1369–1381 (1994a), "Principles determining the structure of β-sheet barrels. I. A theoretical analysis"; Murzin et al. *J Mol Biol 236*:1382–1400 (1994b) "Principles determining the structure of β-sheet barrels. II. The observed structures"; all of these references are explicitly incorporated by reference herein in their entireity).

Four different supersecondary structure parameters useful for α/β proteins are shown in Figure 14. In a preferred embodiment, as for all the supersecondary structure parameters, at least one of these parameter values is altered; other embodiments utilize simultaneous or sequential alteration of two, three or four of these parameter values.

For the α/β protein interactions, the helix center is defined as the average C_α position of the residues chosen for backbone movement. The helix axis is defined as the principal moment of the C_α atoms of these residues (see Chothia et al., 1981, supra). The strand axis is defined as the average of the least-squares lines fit through the midpoints of sequential C_α positions of the two central β-strands. The sheet plane is defined as the least-squares plane fit through the C_α positions of the two central β-strands. The sheet axis is defined as the vector perpendicular to the sheet plane that passes through the helix center. Ω is the angle between the strand axis and the helix axis after projection onto the sheet plane; θ is the angle between the helix axis and the sheet plane; h is the distance between the helix center and the sheet plane; σ is the rotation angle about the helix axis. Backbone alteration requires altering at least one of these parameter values. In a preferred embodiment, the supersecondary structure parameter value Ω is altered by changing the angle degree (either positively or negatively) of up to about 25 degrees, with changes of ± 1°, 2.5°, 5°, 7.5°, and 10° being particularly preferred. In a preferred embodiment, the supersecondary structure parameter value θ is altered by changing the angle degree (either positively or negatively) of up to about 25 degrees, with changes of ± 1°, 2.5°, 5°, 7.5°, and 10° being particularly preferred. the supersecondary structure parameter value σ is altered by changing the angle degree (either positively or negatively) of up to about 25 degrees, with changes of ± 1°, 2.5°, 5°, 7.5°, and 10° being particularly preferred. In a preferred embodiment, the supersecondary structure parameter value h is altered by changes (either positive or negative) of up to about 8 Å, with changes of ± 0.25, 0.50, 0.75, 1.00, 1.25 and 1.5 being particularly preferred. However, as will be appreciated by those in the art, as for all the parameter values outlined herein, larger changes can be made, depending on the protein (i.e. how close or far other secondary structure elements are) and whether other parameter values are made; for example, larger changes in Ω can be made if the helix is also moved away from the sheet (i.e. h is increased).

Four different supersecondary structure parameters useful for α/α proteins are shown in Figure 18. As for α/β parameters, the helix center is defined as the average C_α position of the residues in the helix. The helix axis is defined as the principal moment of the C_α atoms of the residues in the helix. σ is defined as the rotation around the helix axis. Ω is the angle between two strand axes after projection onto a plane. Thus, d, the distance between the helices, can be altered, generally as outlined above for h. Similarly, θ, σ and Ω can be altered as above.

There are a number of different supersecondary structure parameters useful for β/β proteins. β-barrel configurations contain a number of different parameters that can be altered, as shown in Figure 17. These include: (see Figure 17A) R, the barrel radius; α, the angle of tilt of the strands relative to the barrel axis; and b, the interstrand distance; (see Figure 17B) θ, the mean twist of the β-sheet about an axis perpendicular to the strand direction; τ, the mean twist of the β-sheet about an axis parallel to the

strand direction; ε the mean coiling of the β-sheet along the strands; η, the mean coiling of the β-sheet along a line perpendicular to the strands; and (Figure 17C) Ω is angle between the two β-sheet axes. As for the α/β parameter values, each of these may be altered, either positively or negatively. Generally, changes are made in at least one of these parameter values, by changing the angle degree (either positively or negatively) of up to about 25 degrees, with changes of ± 1°, 2.5°, 5°, 7.5°, and 10° being particularly preferred. b can be changed up to ± 1 ·. For β sandwich structures (Figure 17C and 17D), Ω can be altered up to ± 45°, with changes of ± 1°, 2.5°, 5°, 7.5°, and 10° being particularly preferred. Similarly, h can be altered as outlined above for α/β proteins, and θ and φ can be altered up to ± 30°.

Once the desired value changes are selected, the coordinate positions for the positions chosen are altered to reflect the change, to form a "new" or "altered" backbone protein structure, i.e. one that has all or part of the backbone atoms in a different position relative to the starting structure. It should be noted that this process can be repeated, i.e. additional backbone changes can be made, on the same or different residues. In addition, after optimization, the backbone of one or more optimal sequences can altered and an optimization can be run.

Alternatively, movement of the backbone can be done manually, i.e. sections of the protein backbone can be randomly or arbitrarily moved. In this embodiment, the backbone atoms of one or more amino acids can be moved some distance, generally an angstrom or more, in any direction. This can be done using standard modeling programs; for example, Molecular Dynamics minimization, Monte Carlo dynamics, or random backbone coordinate/angle motion. It is also possible to move the backbone atoms of single residues, that are either components of secondary structural elements or not.

Once a protein structure backbone is generated (with alterations, as outlined above) and input into the computer, explicit hydrogens are added if not included within the structure (for example, if the structure was generated by X-ray crystallography, hydrogens must be added). After hydrogen addition, energy minimization of the structure is run, to relax the hydrogens as well as the other atoms, bond angles and bond lengths. In a preferred embodiment, this is done by doing a number of steps of conjugate gradient minimization (Mayo *et al.*, J. Phys. Chem. **94**:8897 (1990)) of atomic coordinate positions to minimize the Dreiding force field with no electrostatics. Generally from about 10 to about 250 steps is preferred, with about 50 being most preferred.

The protein backbone structure contains at least one variable residue position. As is known in the art, the residues, or amino acids, of proteins are generally sequentially numbered starting with the N-terminus of the protein. Thus a protein having a methionine at it's N-terminus is said to have a methionine at residue or amino acid position 1, with the next residues as 2, 3, 4, etc. At each position, the wild type (i.e. naturally occuring) protein may have one of at least 20 amino acids, in any number of rotamers. By "variable residue position" herein is meant an amino acid position of the protein to be designed that is not fixed in the design method as a specific residue or rotamer, generally the wild-type residue or rotamer.

In a preferred embodiment, all of the residue positions of the protein are variable. That is, every amino acid side chain may be altered in the methods of the present invention. This is particularly desirable for smaller proteins, although the present methods allow the design of larger proteins as well. While there is no theoretical limit to the length of the protein which may be designed this way, there is a practical computational limit.

In an alternate preferred embodiment, only some of the residue positions of the protein are variable, and the remainder are "fixed", that is, they are identified in the three dimensional structure as being in a set conformation. In some embodiments, a fixed position is left in its original conformation (which may or may not correlate to a specific rotamer of the rotamer library being used). Alternatively, residues may be fixed as a non-wild type residue; for example, when known site-directed mutagenesis techniques have shown that a particular residue is desirable (for example, to eliminate a proteolytic site or alter the substrate specificity of an enzyme), the residue may be fixed as a particular amino acid. Alternatively, the methods of the present invention may be used to evaluate mutations de novo, as is discussed below. In an alternate preferred embodiment, a fixed position may be "floated"; the amino acid at that position is fixed, but different rotamers of that amino acid are tested. In this embodiment, the variable residues may be at least one, or anywhere from 0.1% to 99.9% of the total number of residues. Thus, for example, it may be possible to change only a few (or one) residues, or most of the residues, with all possibilities in between.

In a preferred embodiment, residues which can be fixed include, but are not limited to, structurally or biologically functional residues. For example, residues which are known to be important for biological activity, such as the residues which form the active site of an enzyme, the substrate binding site of an enzyme, the binding site for a binding partner (ligand/receptor, antigen/antibody, etc.), phosphorylation or glycosylation sites which are crucial to biological function, or structurally important residues, such as disulfide bridges, metal binding sites, critical hydrogen bonding residues, residues critical for backbone conformation such as proline or glycine, residues critical for packing interactions, etc. may all be fixed in a conformation or as a single rotamer, or "floated".

Similarly, residues which may be chosen as variable residues may be those that confer undesirable biological attributes, such as susceptibility to proteolytic degradation, dimerization or aggregation sites, glycosylation sites which may lead to immune responses, unwanted binding activity, unwanted allostery, undesirable enzyme activity but with a preservation of binding, etc.

As will be appreciated by those in the art, the methods of the present invention allow computational testing of "site-directed mutagenesis" targets without actually making the mutants, or prior to making the mutants. That is, quick analysis of sequences in which a small number of residues are changed can be done to evaluate whether a proposed change is desirable. In addition, this may be done on a known protein, or on an protein optimized as described herein.

As will be appreciated by those in the art, a domain of a larger protein may essentially be treated as a small independent protein; that is, a structural or functional domain of a large protein may have

minimal interactions with the remainder of the protein and may essentially be treated as if it were autonomous. In this embodiment, all or part of the residues of the domain may be variable.

It should be noted that even if a position is chosen as a variable position, it is possible that the methods of the invention will optimize the sequence in such a way as to select the wild type residue at the variable position. This generally occurs more frequently for core residues, and less regularly for surface residues. In addition, it is possible to fix residues as non-wild type amino acids as well.

Once the protein backbone structure has been selected and input, and the variable residue positions chosen, a group of potential rotamers for each of the variable residue positions is established. This operation is shown as step 52 in Figure 2. This step may be implemented using the side chain module 32. In one embodiment of the invention, the side chain module 32 includes at least one rotamer library, as described below, and program code that correlates the selected protein backbone structure with corresponding information in the rotamer library. Alternatively, the side chain module 32 may be omitted and the potential rotamers 42 for the selected protein backbone structure may be downloaded through the input/output devices 26.

As is known in the art, each amino acid side chain has a set of possible conformers, called rotamers. See Ponder, *et al.*, Acad. Press Inc. (London) Ltd. pp. 775-791 (1987); Dunbrack, *et al.*, Struc. Biol. **1**(5):334-340 (1994); Desmet, *et al.*, Nature **356**:539-542 (1992), all of which are hereby expressly incorporated by reference in their entireity. Thus, a set of discrete rotamers for every amino acid side chain is used. There are two general types of rotamer libraries: backbone dependent and backbone independent. A backbone dependent rotamer library allows different rotamers depending on the position of the residue in the backbone; thus for example, certain leucine rotamers are allowed if the position is within an α helix, and different leucine rotamers are allowed if the position is not in a α-helix. A backbone independent rotamer library utilizes all rotamers of an amino acid at every position. In general, a backbone independent library is preferred in the consideration of core residues, since flexibility in the core is important. However, backbone independent libraries are computationally more expensive, and thus for surface and boundary positions, a backbone dependent library is preferred. However, either type of library can be used at any position.

In addition, a preferred embodiment does a type of "fine tuning" of the rotamer library by expanding the possible χ (chi) angle values of the rotamers by plus and minus one standard deviation (or more) about the mean value, in order to minimize possible errors that might arise from the discreteness of the library. This is particularly important for aromatic residues, and fairly important for hydrophobic residues, due to the increased requirements for flexibility in the core and the rigidity of aromatic rings; it is not as important for the other residues. Thus a preferred embodiment expands the χ_1 and χ_2 angles for all amino acids except Met, Arg and Lys.

To roughly illustrate the numbers of rotamers, in one version of the Dunbrack & Karplus backbone-dependent rotamer library, alanine has 1 rotamer, glycine has 1 rotamer, arginine has 55 rotamers, threonine has 9 rotamers, lysine has 57 rotamers, glutamic acid has 69 rotamers, asparagine has 54

rotamers, aspartic acid has 27 rotamers, tryptophan has 54 rotamers, tyrosine has 36 rotamers, cysteine has 9 rotamers, glutamine has 69 rotamers, histidine has 54 rotamers, valine has 9 rotamers, isoleucine has 45 rotamers, leucine has 36 rotamers, methionine has 21 rotamers, serine has 9 rotamers, and phenylalanine has 36 rotamers.

In general, proline is not generally used, since it will rarely be chosen for any position, although it can be included if desired. Similarly, a preferred embodiment omits cysteine as a consideration, only to avoid potential disulfide problems, although it can be included if desired.

As will be appreciated by those in the art, other rotamer libraries with all dihedral angles staggered can be used or generated.

In a preferred embodiment, at a minimum, at least one variable position has rotamers from at least two different amino acid side chains; that is, a sequence is being optimized, rather than a structure.

In a preferred embodiment, rotamers from all of the amino acids (or all of them except cysteine, glycine and proline) are used for each variable residue position; that is, the group or set of potential rotamers at each variable position is every possible rotamer of each amino acid. This is especially preferred when the number of variable positions is not high as this type of analysis can be computationally expensive.

In a preferred embodiment, each variable position is classified as either a core, surface or boundary residue position, although in some cases, as explained below, the variable position may be set to glycine to minimize backbone strain.

It should be understood that quantitative protein design or optimization studies prior to the present invention focused almost exclusively on core residues. The present invention, however, provides methods for designing proteins containing core, surface and boundary positions. Alternate embodiments utilize methods for designing proteins containing core and surface residues, core and boundary residues, and surface and boundary residues, as well as core residues alone (using the scoring functions of the present invention), surface residues alone, or boundary residues alone.

The classification of residue positions as core, surface or boundary may be done in several ways, as will be appreciated by those in the art. In a preferred embodiment, the classification is done via a visual scan of the original protein backbone structure, including the side chains, and assigning a classification based on a subjective evaluation of one skilled in the art of protein modelling. Alternatively, a preferred embodiment utilizes an assessment of the orientation of the Cα-Cβ vectors relative to a solvent accessible surface computed using only the template Cα atoms. In a preferred embodiment, the solvent accessible surface for only the Cα atoms of the target fold is generated using the Connolly algorithm with a probe radius ranging from about 4 to about 12 ·, with from about 6 to about 10 · being preferred, and 8 · being particularly preferred. The Cα radius used ranges from about 1.6 · to about 2.3 ·, with from about 1.8 to about 2.1 · being preferred, and 1.95 · being

especially preferred. A residue is classified as a core position if a) the distance for its Cα, along its Cα-Cβ vector, to the solvent accessible surface is greater than about 4-6 ·, with greater than about 5.0 · being especially preferred, and b) the distance for its Cβ to the nearest surface point is greater than about 1.5-3 ·, with greater than about 2.0 · being especially preferred. The remaining residues are classified as surface positions if the sum of the distances from their Cα, along their Cα-Cβ vector, to the solvent accessible surface, plus the distance from their Cβ to the closest surface point was less than about 2.5-4 ·, with less than about 2.7 · being especially preferred. All remaining residues are classified as boundary positions.

Once each variable position is classified as either core, surface or boundary, a set of amino acid side chains, and thus a set of rotamers, is assigned to each position. That is, the set of possible amino acid side chains that the program will allow to be considered at any particular position is chosen. Subsequently, once the possible amino acid side chains are chosen, the set of rotamers that will be evaluated at a particular position can be determined. Thus, a core residue will generally be selected from the group of hydrophobic residues consisting of alanine, valine, isoleucine, leucine, phenylalanine, tyrosine, tryptophan, and methionine (in some embodiments, when the α scaling factor of the van der Waals scoring function, described below, is low, methionine is removed from the set), and the rotamer set for each core position potentially includes rotamers for these eight amino acid side chains (all the rotamers if a backbone independent library is used, and subsets if a rotamer dependent backbone is used). Similarly, surface positions are generally selected from the group of hydrophilic residues consisting of alanine, serine, threonine, aspartic acid, asparagine, glutamine, glutamic acid, arginine, lysine and histidine. The rotamer set for each surface position thus includes rotamers for these ten residues. Finally, boundary positions are generally chosen from alanine, serine, threonine, aspartic acid, asparagine, glutamine, glutamic acid, arginine, lysine histidine, valine, isoleucine, leucine, phenylalanine, tyrosine, tryptophan, and methionine. The rotamer set for each boundary position thus potentially includes every rotamer for these seventeen residues (assuming cysteine, glycine and proline are not used, although they can be).

Thus, as will be appreciated by those in the art, there is a computational benefit to classifying the residue positions, as it decreases the number of calculations. It should also be noted that there may be situations where the sets of core, boundary and surface residues are altered from those described above; for example, under some circumstances, one or more amino acids is either added or subtracted from the set of allowed amino acids. For example, some proteins which dimerize or multimerize, or have ligand binding sites, may contain hydrophobic surface residues, etc. In addition, residues that do not allow helix "capping" or the favorable interaction with an α-helix dipole may be subtracted from a set of allowed residues. This modification of amino acid groups is done on a residue by residue basis.

In a preferred embodiment, proline, cysteine and glycine are not included in the list of possible amino acid side chains, and thus the rotamers for these side chains are not used. However, in a preferred embodiment, when the variable residue position has a φ angle (that is, the dihedral angle defined by

1) the carbonyl carbon of the preceding amino acid; 2) the nitrogen atom of the current residue; 3) the α-carbon of the current residue; and 4) the carbonyl carbon of the current residue) greater than 0°, the position is set to glycine to minimize backbone strain.

Once the group of potential rotamers is assigned for each variable residue position, processing proceeds to step 54 of Figure 2. This processing step entails analyzing interactions of the rotamers with each other and with the protein backbone to generate optimized protein sequences. The ranking module 34 may be used to perform these operations. That is, computer code is written to implement the following functions. Simplistically, as is generally outlined above, the processing initially comprises the use of a number of scoring functions, described below, to calculate energies of interactions of the rotamers, either to the backbone itself or other rotamers.

The scoring functions include a Van der Waals potential scoring function, a hydrogen bond potential scoring function, an atomic solvation scoring function, a secondary structure propensity scoring function and an electrostatic scoring function. As is further described below, at least one scoring function is used to score each position, although the scoring functions may differ depending on the position classification or other considerations, like favorable interaction with an α-helix dipole. As outlined below, the total energy which is used in the calculations is the sum of the energy of each scoring function used at a particular position, as is generally shown in Equation 1:

Equation 1

$$E_{total} = nE_{vdw} + nE_{as} + nE_{h\text{-}bonding} + nE_{ss} + nE_{elec}$$

In Equation 1, the total energy is the sum of the energy of the van der Waals potential (E_{vdw}), the energy of atomic solvation (E_{as}), the energy of hydrogen bonding ($E_{h\text{-}bonding}$), the energy of secondary structure (E_{ss}) and the energy of electrostatic interaction (E_{elec}). The term n is either 0 or 1, depending on whether the term is to be considered for the particular residue position, as is more fully outlined below.

In a preferred embodiment, a van der Waals' scoring function is used. As is known in the art, van der Waals' forces are the weak, non-covalent and non-ionic forces between atoms and molecules, that is, the induced dipole and electron repulsion (Pauli principle) forces.

The van der Waals scoring function is based on a van der Waals potential energy. There are a number of van der Waals potential energy calculations, including a Lennard-Jones 12/6 potential with radii and well depth parameters from the Dreiding force field, Mayo *et al.*, J. Prot. Chem., 1990, expressly incorporated herein by reference, or the exponential 6 potential. Equation 2, shown below, is the preferred Lennard-Jones potential:

Equation 2

$$E_{vdw} = D_0\left\{\left(\frac{R_0}{R}\right)^{12} - 2\left(\frac{R_0}{R}\right)^{6}\right\}$$

R_0 is the geometric mean of the van der Waals radii of the two atoms under consideration, and D_0 is the geometric mean of the well depth of the two atoms under consideration. E_{vdw} and R are the energy and interatomic distance between the two atoms under consideration, as is more fully described below.

In a preferred embodiment, the van der Waals forces are scaled using a scaling factor, α, as is generally discussed in Example 4. Equation 3 shows the use of α in the van der Waals Lennard-Jones potential equation:

Equation 3

$$E_{vdw} = D_0\left\{\left(\frac{\alpha R_0}{R}\right)^{12} - 2\left(\frac{\alpha R_0}{R}\right)^{6}\right\}$$

The role of the α scaling factor is to change the importance of packing effects in the optimization and design of any particular protein. As discussed in the Examples, different values for α result in different sequences being generated by the present methods. Specifically, a reduced van der Waals steric constraint can compensate for the restrictive effect of a fixed backbone and discrete side-chain rotamers in the simulation and can allow a broader sampling of sequences compatible with a desired fold. In a preferred embodiment, α values ranging from about 0.70 to about 1.10 can be used, with α values from about 0.8 to about 1.05 being preferred, and from about 0.85 to about 1.0 being especially preferred. Specific α values which are preferred are 0.80, 0.85, 0.90, 0.95, 1.00, and 1.05.

Generally speaking, variation of the van der Waals scale factor α results in four regimes of packing specificity: regime 1 where $0.9 \leq \alpha \leq 1.05$ and packing constraints dominate the sequence selection; regime 2 where $0.8 \leq \alpha < 0.9$ and the hydrophobic solvation potential begins to compete with packing forces; regime 3 where $\alpha < 0.8$ and hydrophobic solvation dominates the design; and, regime 4 where $\alpha > 1.05$ and van der Waals repulsions appear to be too severe to allow meaningful sequence selection. In particular, different α values may be used for core, surface and boundary positions, with regimes 1 and 2 being preferred for core residues, regime 1 being preferred for surface residues, and regime 1 and 2 being preferred for boundary residues.

In a preferred embodiment, the van der Waals scaling factor is used in the total energy calculations for each variable residue position, including core, surface and boundary positions.

In a preferred embodiment, an atomic solvation potential scoring function is used. As is appreciated by those in the art, solvent interactions of a protein are a significant factor in protein stability, and residue/protein hydrophobicity has been shown to be the major driving force in protein folding. Thus, there is an entropic cost to solvating hydrophobic surfaces, in addition to the potential for misfolding or aggregation. Accordingly, the burial of hydrophobic surfaces within a protein structure is beneficial to both folding and stability. Similarly, there can be a disadvantage for burying hydrophilic residues. The accessible surface area of a protein atom is generally defined as the area of the surface over

which a water molecule can be placed while making van der Waals contact with this atom and not penetrating any other protein atom. Thus, in a preferred embodiment, the solvation potential is generally scored by taking the total possible exposed surface area of the moiety or two independent moieties (either a rotamer or the first rotamer and the second rotamer), which is the reference, and subtracting out the "buried" area, i.e. the area which is not solvent exposed due to interactions either with the backbone or with other rotamers. This thus gives the exposed surface area.

Alternatively, a preferred embodiment calculates the scoring function on the basis of the "buried" portion; i.e. the total possible exposed surface area is calculated, and then the calculated surface area after the interaction of the moieties is subtracted, leaving the buried surface area. A particularly preferred method does both of these calculations.

As is more fully described below, both of these methods can be done in a variety of ways. See Eisenberg *et al.*, Nature **319**:199-203 (1986); Connolly, Science **221**:709-713 (1983); and Wodak, *et al.*, Proc. Natl. Acad. Sci. USA **77**(4):1736-1740 (1980), all of which are expressly incorporated herein by reference. As will be appreciated by those in the art, this solvation potential scoring function is conformation dependent, rather than conformation independent.

In a preferred embodiment, the pairwise solvation potential is implemented in two components, "singles" (rotamer/template) and "doubles" (rotamer/rotamer), as is more fully described below. For the rotamer/template buried area, the reference state is defined as the rotamer in question at residue position i with the backbone atoms only of residues i-1, i and i+1, although in some instances just i may be used. Thus, in a preferred embodiment, the solvation potential is not calculated for the interaction of each backbone atom with a particular rotamer, although more may be done as required. The area of the side chain is calculated with the backbone atoms excluding solvent but not counted in the area. The folded state is defined as the area of the rotamer in question at residue i, but now in the context of the entire template structure including non-optimized side chains, i.e. every other foxed position residue. The rotamer/template buried area is the difference between the reference and the folded states. The rotamer/rotamer reference area can be done in two ways; one by using simply the sum of the areas of the isolated rotamers; the second includes the full backbone. The folded state is the area of the two rotamers placed in their relative positions on the protein scaffold but with no template atoms present. In a preferred embodiment, the Richards definition of solvent accessible surface area (Lee and Richards, J. Mol. Biol. 55:379-400, 1971, hereby incorporated by reference) is used, with a probe radius ranging from 0.8 to 1.6 Å, with 1.4 Å being preferred, and Drieding van der Waals radii, scaled from 0.8 to 1.0. Carbon and sulfur, and all attached hydrogens, are considered nonpolar. Nitrogen and oxygen, and all attached hydrogens, are considered polar. Surface areas are calculated with the Connolly algorithm using a dot density of 10 Å-2 (Connolly, (1983) (supra), hereby incorporated by reference).

In a preferred embodiment, there is a correction for a possible overestimation of buried surface area which may exist in the calculation of the energy of interaction between two rotamers (but not the interaction of a rotamer with the backbone). Since, as is generally outlined below, rotamers are only

considered in pairs, that is, a first rotamer is only compared to a second rotamer during the "doubles" calculations, this may overestimate the amount of buried surface area in locations where more than two rotamers interact, that is, where rotamers from three or more residue positions come together. Thus, a correction or scaling factor is used as outlined below.

The general energy of solvation is shown in Equation 4:

Equation 4

$$E_{sa} = f(SA)$$

where E_{sa} is the energy of solvation, f is a constant used to correlate surface area and energy, and SA is the surface area. This equation can be broken down, depending on which parameter is being evaluated. Thus, when the hydrophobic buried surface area is used, Equation 5 is appropriate:

Equation 5

$$E_{sa} = f_1(SA_{\text{buried hydrophobic}})$$

where f_1 is a constant which ranges from about 10 to about 50 cal/mol/ $Å^2$, with 23 or 26 cal/mol/ $Å^2$ being preferred. When a penalty for hydrophilic burial is being considered, the equation is shown in Equation 6:

Equation 6

$$E_{sa} = f_1(SA_{\text{buried hydrophobic}}) + f_2(SA_{\text{buried hydrophilic}})$$

where f_2 is a constant which ranges from -50 to -250 cal/mol/ $Å^2$, with -86 or -100 cal/mol/ $Å^2$ being preferred. Similarly, if a penalty for hydrophobic exposure is used, equation 7 or 8 may be used:

Equation 7

$$E_{sa} = f_1(SA_{\text{buried hydrophobic}}) + f_3(SA_{\text{exposed hydrophobic}})$$

Equation 8

$$E_{sa} = f_1(SA_{\text{buried hydrophobic}}) + f_2(SA_{\text{buried hydrophilic}}) + f_3(SA_{\text{exposed hydrophobic}}) + f_4(SA_{\text{exposed hydrophilic}})$$

In a preferred embodiment, $f_3 = -f_1$.

In one embodiment, backbone atoms are not included in the calculation of surface areas, and values of 23 cal/mol/ $Å^2$ (f_1) and -86 cal/mol/ $Å^2$ (f_2) are determined.

In a preferred embodiment, this overcounting problem is addressed using a scaling factor that compensates for only the portion of the expression for pairwise area that is subject to over-counting. In this embodiment, values of -26 cal/mol/ $Å^2$ (f_1) and 100 cal/mol/ $Å^2$ (f_2) are determined.

Atomic solvation energy is expensive, in terms of computational time and resources. Accordingly, in a preferred embodiment, the solvation energy is calculated for core and/or boundary residues, but not

surface residues, with both a calculation for core and boundary residues being preferred, although any combination of the three is possible.

In a preferred embodiment, a hydrogen bond potential scoring function is used. A hydrogen bond potential is used as predicted hydrogen bonds do contribute to designed protein stability (see Stickle *et al.*, J. Mol. Biol. 226:1143 (1992); Huyghues-Despointes *et al.*, Biochem. 34:13267 (1995), both of which are expressly incorporated herein by reference). As outlined previously, explicit hydrogens are generated on the protein backbone structure.

In a preferred embodiment, the hydrogen bond potential consists of a distance-dependent term and an angle-dependent term, as shown in Equation 9:

Equation 9

$$E_{\text{H-Bonding}} = D_0 \left\{ 5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10} \right\} F(\theta, \text{Ø}, \varphi)$$

where R_0 (2.8 Å) and D_0 (8 kcal/mol) are the hydrogen-bond equilibrium distance and well-depth, respectively, and R is the donor to acceptor distance. This hydrogen bond potential is based on the potential used in DREIDING with more restrictive angle-dependent terms to limit the occurrence of unfavorable hydrogen bond geometries. The angle term varies depending on the hybridization state of the donor and acceptor, as shown in Equations 10, 11, 12 and 13. Equation 10 is used for sp^3 donor to sp^3 acceptor; Equation 11 is used for sp^3 donor to sp^2 acceptor, Equation 12 is used for sp^2 donor to sp^3 acceptor, and Equation 13 is used for sp^2 donor to sp^2 acceptor:

Equation 10

$$F = \cos^2\theta \cos^2(\text{Ø} - 109.5)$$

Equation 11

$$F = \cos^2\theta \cos^2 \text{Ø}$$

Equation 12

$$F = \cos^4\theta$$

Equation 13

$$F = \cos^2\theta \cos^2(\max[\varphi, \varphi])$$

In Equations 10-13, θ is the donor-hydrogen-acceptor angle, φ is the hydrogen-acceptor-base angle (the base is the atom attached to the acceptor, for example the carbonyl carbon is the base for a carbonyl oxygen acceptor), and φ is the angle between the normals of the planes defined by the six atoms attached to the sp^2 centers (the supplement of φ is used when φ is less than 90°). The hydrogen-bond function is only evaluated when 2.6 Å ≤ R ≤ 3.2 Å, θ > 90°, φ - 109.5° < 90° for the sp^3

donor - sp^3 acceptor case, and, $\varphi > 90°$ for the sp^3 donor - sp^2 acceptor case; preferably, no switching functions are used. Template donors and acceptors that are involved in template-template hydrogen bonds are preferably not included in the donor and acceptor lists. For the purpose of exclusion, a template-template hydrogen bond is considered to exist when 2.5 Å ≤ R ≤ 3.3 Å and θ ≥ 135°.

The hydrogen-bond potential may also be combined or used with a weak coulombic term that includes a distance-dependent dielectric constant of 40R, where R is the interatomic distance. Partial atomic charges are preferably only applied to polar functional groups. A net formal charge of +1 is used for Arg and Lys and a net formal charge of -1 is used for Asp and Glu; see Gasteiger, *et al.*, Tetrahedron **36**:3219-3288 (1980); Rappe, *et al.*, J. Phys. Chem. **95**:3358-3363 (1991).

In a preferred embodiment, an explicit penalty is given for buried polar hydrogen atoms which are not hydrogen bonded to another atom. See Eisenberg, *et al.*, (1986) (supra), hereby expressly incorporated by reference. In a preferred embodiment, this penalty for polar hydrogen burial, is from about 0 to about 3 kcal/mol, with from about 1 to about 3 being preferred and 2 kcal/mol being particularly preferred. This penalty is only applied to buried polar hydrogens not involved in hydrogen bonds. A hydrogen bond is considered to exist when E_{HB} ranges from about 1 to about 4 kcal/mol, with E_{HB} of less than -2 kcal/mol being preferred. In addition, in a preferred embodiment, the penalty is not applied to template hydrogens, i.e. unpaired buried hydrogens of the backbone.

In a preferred embodiment, only hydrogen bonds between a first rotamer and the backbone are scored, and rotamer-rotamer hydrogen bonds are not scored. In an alternative embodiment, hydrogen bonds between a first rotamer and the backbone are scored, and rotamer-rotamer hydrogen bonds are scaled by 0.5.

In a preferred embodiment, the hydrogen bonding scoring function is used for all positions, including core, surface and boundary positions. In alternate embodiments, the hydrogen bonding scoring function may be used on only one or two of these.

In a preferred embodiment, a secondary structure propensity scoring function is used. This is based on the specific amino acid side chain, and is conformation independent. That is, each amino acid has a certain propensity to take on a secondary structure, either α-helix or β-sheet, based on its φ and ψ angles. See Muñoz *et al.*, Current Op. in Biotech. **6**:382 (1995); Minor, *et al.*, Nature **367**:660-663 (1994); Padmanabhan, *et al.*, Nature **344**:268-270 (1990); Muñoz, *et al.*, Folding & Design **1**(3):167-178 (1996); and Chakrabartty, *et al.*, Protein Sci. 3:843 (1994), all of which are expressly incorporated herein by reference. Thus, for variable residue positions that are in recognizable secondary structure in the backbone, a secondary structure propensity scoring function is preferably used. That is, when a variable residue position is in an α-helical area of the backbone, the α-helical propensity scoring function described below is calculated. Whether or not a position is in a α-helical area of the backbone is determined as will be appreciated by those in the art, generally on the basis of φ and ψ angles; for α-helix, φ angles from -2 to -70 and ψ angles from -30 to -100 generally describe an α-helical area of the backbone.

Similarly, when a variable residue position is in a β-sheet backbone conformation, the β-sheet propensity scoring function is used. β-sheet backbone conformation is generally described by φ angles from -30 to -100 and χ angles from +40 to +180. In alternate preferred embodiments, variable residue positions which are within areas of the backbone which are not assignable to either β-sheet or α-helix structure may also be subjected to secondary structure propensity calculations.

In a preferred embodiment, energies associated with secondary propensities are calculated using Equation 14:

Equation 14

$$E_\alpha = 10^{N_{ss}(\Delta G^\circ_{aa} - \Delta G^\circ_{ala})} - 1$$

In Equation 14, E_α (or Eβ) is the energy of α-helical propensity, ΔG°_{aa} is the standard free energy of helix propagation of the amino acid, and ΔG°_{ala} is the standard free energy of helix propagation of alanine used as a standard, or standard free energy of β-sheet formation of the amino acid, both of which are available in the literature (see Chakrabartty, *et al.*, (1994) (supra), and Munoz, *et al.*, Folding & Design 1(3):167-178 (1996)), both of which are expressly incorporated herein by reference), and N_{ss} is the propensity scale factor which is set to range from 1 to 4, with 3.0 being preferred. This potential is preferably selected in order to scale the propensity energies to a similar range as the other terms in the scoring function.

In a preferred embodiment, β-sheet propensities are preferably calculated only where the i-1 and i+1 residues are also in β-sheet conformation.

In a preferred embodiment, the secondary structure propensity scoring function is used only in the energy calculations for surface variable residue positions. In alternate embodiments, the secondary structure propensity scoring function is used in the calculations for core and boundary regions as well.

In a preferred embodiment, an electrostatic scoring function is used, as shown below in Equation 15:

Equation 15

$$E_{elec} = \frac{qq'}{\varepsilon r^2}$$

In this Equation, q is the charge on atom 1, q' is charge on atom 2, and r is the interaction distance.

In a preferred embodiment, at least one scoring function is used for each variable residue position; in preferred embodiments, two, three or four scoring functions are used for each variable residue position.

Once the scoring functions to be used are identified for each variable position, the preferred first step in the computational analysis comprises the determination of the interaction of each possible rotamer with all or part of the remainder of the protein. That is, the energy of interaction, as measured by one

or more of the scoring functions, of each possible rotamer at each variable residue position with either the backbone or other rotamers, is calculated. In a preferred embodiment, the interaction of each rotamer with the entire remainder of the protein, i.e. both the entire template and all other rotamers, is done. However, as outlined above, it is possible to only model a portion of a protein, for example a domain of a larger protein, and thus in some cases, not all of the protein need be considered.

In a preferred embodiment, the first step of the computational processing is done by calculating two sets of interactions for each rotamer at every position (step 70 of figure 3): the interaction of the rotamer side chain with the template or backbone (the "singles" energy), and the interaction of the rotamer side chain with all other possible rotamers at every other position (the "doubles" energy), whether that position is varied or floated. It should be understood that the backbone in this case includes both the atoms of the protein structure backbone, as well as the atoms of any fixed residues, wherein the fixed residues are defined as a particular conformation of an amino acid.

Thus, "singles" (rotamer/template) energies are calculated for the interaction of every possible rotamer at every variable residue position with the backbone, using some or all of the scoring functions. Thus, for the hydrogen bonding scoring function, every hydrogen bonding atom of the rotamer and every hydrogen bonding atom of the backbone is evaluated, and the E_{HB} is calculated for each possible rotamer at every variable position. Similarly, for the van der Waals scoring function, every atom of the rotamer is compared to every atom of the template (generally excluding the backbone atoms of its own residue), and the E_{vdW} is calculated for each possible rotamer at every variable residue position. In addition, generally no van der Waals energy is calculated if the atoms are connected by three bonds or less. For the atomic solvation scoring function, the surface of the rotamer is measured against the surface of the template, and the E_{as} for each possible rotamer at every variable residue position is calculated. The secondary structure propensity scoring function is also considered as a singles energy, and thus the total singles energy may contain an E_{ss} term. As will be appreciated by those in the art, many of these energy terms will be close to zero, depending on the physical distance between the rotamer and the template position; that is, the farther apart the two moieties, the lower the energy.

Accordingly, as outlined above, the total singles energy is the sum of the energy of each scoring function used at a particular position, as shown in Equation 1, wherein n is either 1 or zero, depending on whether that particular scoring function was used at the rotamer position:

Equation 1

$$E_{total} = nE_{vdw} + nE_{as} + nE_{h\text{-}bonding} + nE_{ss} + nE_{elec}$$

Once calculated, each singles E_{total} for each possible rotamer is stored in the memory 24 within the computer, such that it may be used in subsequent calculations, as outlined below.

For the calculation of "doubles" energy (rotamer/rotamer), the interaction energy of each possible rotamer is compared with every possible rotamer at all other variable residue positions. Thus,

"doubles" energies are calculated for the interaction of every possible rotamer at every variable residue position with every possible rotamer at every other variable residue position, using some or all of the scoring functions. Thus, for the hydrogen bonding scoring function, every hydrogen bonding atom of the first rotamer and every hydrogen bonding atom of every possible second rotamer is evaluated, and the E_{HB} is calculated for each possible rotamer pair for any two variable positions. Similarly, for the van der Waals scoring function, every atom of the first rotamer is compared to every atom of every possible second rotamer, and the E_{vdW} is calculated for each possible rotamer pair at every two variable residue positions. For the atomic solvation scoring function, the surface of the first rotamer is measured against the surface of every possible second rotamer, and the E_{as} for each possible rotamer pair at every two variable residue positions is calculated. The secondary structure propensity scoring function need not be run as a "doubles" energy, as it is considered as a component of the "singles" energy. As will be appreciated by those in the art, many of these double energy terms will be close to zero, depending on the physical distance between the first rotamer and the second rotamer; that is, the farther apart the two moieties, the lower the energy.

Accordingly, as outlined above, the total doubles energy is the sum of the energy of each scoring function used to evaluate every possible pair of rotamers, as shown in Equation 16, wherein n is either 1 or zero, depending on whether that particular scoring function was used at the rotamer position:

Equation 16

$$E_{total} = nE_{vdw} + nE_{as} + nE_{h\text{-}bonding} + E_{elec}$$

An example is illuminating. A first variable position, i, has three (an unrealistically low number) possible rotamers (which may be either from a single amino acid or different amino acids) which are labelled ia, ib, and ic. A second variable position, j, also has three possible rotamers, labelled jd, je, and jf. Thus, nine doubles energies (E_{total}) are calculated in all: E_{total}(ia, jd), E_{total}(ia, je), E_{total}(ia, jf), E_{total}(ib, jd), E_{total}(ib, je), E_{total}(ib, jf), E_{total}(ic, jd), E_{total}(ic, je), and E_{total}(ic, jf).

Once calculated, each doubles E_{total} for each possible rotamer pair is stored in memory 24 within the computer, such that it may be used in subsequent calculations, as outlined below.

Once the singles and doubles energies are calculated and stored, the next step of the computational processing may occur. Generally speaking, the goal of the computational processing is to determine a set of optimized protein sequences. By "optimized protein sequence" herein is meant a sequence that best fits the mathematical equations herein. As will be appreciated by those in the art, a global optimized sequence is the one sequence that best fits Equation 1, i.e. the sequence that has the lowest energy of any possible sequence. However, there are any number of sequences that are not the global minimum but that have low energies.

In a preferred embodiment, the set comprises the globally optimal sequence in its optimal conformation, i.e. the optimum rotamer at each variable position. That is, computational processing is run until the simulation program converges on a single sequence which is the global optimum.

In a preferred embodiment, the set comprises at least two optimized protein sequences. Thus for example, the computational processing step may eliminate a number of disfavored combinations but be stopped prior to convergence, providing a set of sequences of which the global optimum is one. In addition, further computational analysis, for example using a different method, may be run on the set, to further eliminate sequences or rank them differently. Alternatively, as is more fully described below, the global optimum may be reached, and then further computational processing may occur, which generates additional optimized sequences in the neighborhood of the global optimum.

If a set comprising more than one optimized protein sequences is generated, they may be rank ordered in terms of theoretical quantitative stability, as is more fully described below.

In a preferred embodiment, the computational processing step first comprises an elimination step, sometimes referred to as "applying a cutoff", either a singles elimination or a doubles elimination. Singles elimination comprises the elimination of all rotamers with template interaction energies of greater than about 10 kcal/mol prior to any computation, with elimination energies of greater than about 15 kcal/mol being preferred and greater than about 25 kcal/mol being especially preferred. Similarly, doubles elimination is done when a rotamer has interaction energies greater than about 10 kcal/mol with all rotamers at a second residue position, with energies greater than about 15 being preferred and greater than about 25 kcal/mol being especially preferred.

In a preferred embodiment, the computational processing comprises direct determination of total sequence energies, followed by comparison of the total sequence energies to ascertain the global optimum and rank order the other possible sequences, if desired. The energy of a total sequence is shown below in Equation 17:

Equation 17

$$E_{\text{total protein}} = E_{(b\text{-}b)} + \sum_{\text{all i}} E_{(ia)} + \sum_{\text{all i}} \sum_{\text{all j pairs}} E_{(ia,\,ja)}$$

Thus every possible combination of rotamers may be directly evaluated by adding the backbone-backbone (sometimes referred to herein as template-template) energy ($E_{(b\text{-}b)}$ which is constant over all sequences herein since the backbone is kept constant), the singles energy for each rotamer (which has already been calculated and stored), and the doubles energy for each rotamer pair (which has already been calculated and stored). Each total sequence energy of each possible rotamer sequence can then be ranked, either from best to worst or worst to best. This is obviously computationally expensive and becomes unwieldy as the length of the protein increases.

In a preferred embodiment, the computational processing includes one or more Dead-End Elimination (DEE) computational steps. The DEE theorem is the basis for a very fast discrete search program that was designed to pack protein side chains on a fixed backbone with a known sequence. See Desmet, *et al.*, Nature **356**:539-542 (1992); Desmet, *et al.*, The Protein Folding Problem and Tertiary Structure Prediction, Ch. **10**:1-49 (1994); Goldstein, Biophys. Jour. **66**:1335-1340 (1994), all of which are incorporated herein by reference. DEE is based on the observation that if a rotamer can be eliminated from consideration at a particular position, i.e. make a determination that a particular rotamer is definitely not part of the global optimal conformation, the size of the search is reduced. This is done by comparing the worst interaction (i.e. energy or E_{total}) of a first rotamer at a single variable position with the best interaction of a second rotamer at the same variable position. If the worst interaction of the first rotamer is still better than the best interaction of the second rotamer, then the second rotamer cannot possibly be in the optimal conformation of the sequence. The original DEE theorem is shown in Equation 18:

Equation 18

$$E(ia) + \sum_j [\min \text{ over } t\{E(ia, jt)\}] > E(ib) + \sum_j [\max \text{ over } t\{E(ib, jt)\}]$$

In Equation 18, rotamer ia is being compared to rotamer ib. The left side of the inequality is the best possible interaction energy (E_{total}) of ia with the rest of the protein; that is, "min over t" means find the rotamer t on position j that has the best interaction with rotamer ia. Similarly, the right side of the inequality is the worst possible (max) interaction energy of rotamer ib with the rest of the protein. If this inequality is true, then rotamer ia is Dead-Ending and can be Eliminated. The speed of DEE comes from the fact that the theorem only requires sums over the sequence length to test and eliminate rotamers.

In a preferred embodiment, a variation of DEE is performed. Goldstein DEE, based on Goldstein, (1994) (supra), hereby expressly incorporated by reference, is a variation of the DEE computation, as shown in Equation 19:

Equation 19

$$E(ia) - E(ib) + \sum_j [\min \text{ over } t\{E(ia, jt) - E(ib, jt)\}] > 0$$

In essence, the Goldstein Equation 19 says that a first rotamer a of a particular position i (rotamer ia) will not contribute to a local energy minimum if the energy of conformation with ia can always be lowered by just changing the rotamer at that position to ib, keeping the other residues equal. If this inequality is true, then rotamer ia is Dead-Ending and can be Eliminated.

Thus, in a preferred embodiment, a first DEE computation is done where rotamers at a single variable position are compared, ("singles" DEE) to eliminate rotamers at a single position. This analysis is repeated for every variable position, to eliminate as many single rotamers as possible. In addition, every time a rotamer is eliminated from consideration through DEE, the minimum and maximum

calculations of Equation 18 or 19 change, depending on which DEE variation is used, thus conceivably allowing the elimination of further rotamers. Accordingly, the singles DEE computation can be repeated until no more rotamers can be eliminated; that is, when the inequality is not longer true such that all of them could conceivably be found on the global optimum.

In a preferred embodiment, "doubles" DEE is additionally done. In doubles DEE, pairs of rotamers are evaluated; that is, a first rotamer at a first position and a second rotamer at a second position are compared to a third rotamer at the first position and a fourth rotamer at the second position, either using original or Goldstein DEE. Pairs are then flagged as nonallowable, although single rotamers cannot be eliminated, only the pair. Again, as for singles DEE, every time a rotamer pair is flagged as nonallowable, the minimum calculations of Equation 18 or 19 change (depending on which DEE variation is used) thus conceivably allowing the flagging of further rotamer pairs. Accordingly, the doubles DEE computation can be repeated until no more rotamer pairs can be flagged; that is, where the energy of rotamer pairs overlap such that all of them could conceivably be found on the global optimum.

In addition, in a preferred embodiment, rotamer pairs are initially prescreened to eliminate rotamer pairs prior to DEE. This is done by doing relatively computationally inexpensive calculations to eliminate certain pairs up front. This may be done in several ways, as is outlined below.

In a preferred embodiment, the rotamer pair with the lowest interaction energy with the rest of the system is found. Inspection of the energy distributions in sample matrices has revealed that an $i_w j_v$ pair that dead-end eliminates a particular $i_r j_s$ pair can also eliminate other $i_r j_s$ pairs. In fact, there are often a few $i_w j_v$ pairs, which we call "magic bullets," that eliminate a significant number of $i_r j_s$ pairs. We have found that one of the most potent magic bullets is the pair for which maximum interaction energy, $t_{max}([i_w j_v])k_t$, is least. This pair is referred to as $[i_w j_v]_{mb}$. If this rotamer pair is used in the first round of doubles DEE, it tends to eliminate pairs faster.

Our first speed enhancement is to evaluate the first-order doubles calculation for only the matrix elements in the row corresponding to the $[i_w j_v]_{mb}$ pair. The discovery of $[i_w j_v]_{mb}$ is an n^2 calculation (n = the number of rotamers per position), and the application of Equation 19 to the single row of the matrix corresponding to this rotamer pair is another n^2 calculation, so the calculation time is small in comparison to a full first-order doubles calculation. In practice, this calculation produces a large number of dead-ending pairs, often enough to proceed to the next iteration of singles elimination without any further searching of the doubles matrix.

The magic bullet first-order calculation will also discover all dead-ending pairs that would be discovered by the Equation 18 or 19, thereby making it unnecessary. This stems from the fact that $\varepsilon_{max}([i_w j_v]_{mb})$ must be less than or equal to any $\varepsilon_{max}([i_w j_v])$ that would successfully eliminate a pair by he Equation 18 or 19.

Since the minima and maxima of any given pair has been precalculated as outlined herein, a second speed-enhancement precalculation may be done. By comparing extrema, pairs that will not dead end can be identified and thus skipped, reducing the time of the DEE calculation. Thus, pairs that satisfy either one of the following criteria are skipped:

Equation 20

$$\varepsilon_{\min}([i_r j_s]) < \varepsilon_{\min}([i_u j_v])$$

Equation 21:

$$\varepsilon_{\max}([i_r j_s]) < \varepsilon_{\max}([i_u j_v])$$

Because the matrix containing these calculations is symmetrical, half of its elements will satisfy the first inequality Equation 20, and half of those remaining will satisfy the other inequality Equation 21. These three quarters of the matrix need not be subjected to the evaluation of Equation 18 or 19, resulting in a theoretical speed enhancement of a factor of four.

The last DEE speed enhancement refines the search of the remaining quarter of the matrix. This is done by constructing a metric from the precomputed extrema to detect those matrix elements likely to result in a dead-ending pair.

A metric was found through analysis of matrices from different sample optimizations. We searched for combinations of the extrema that predicted the likelihood that a matrix element would produce a dead-ending pair. Interval sizes (see Figure 12) for each pair were computed from differences of the extrema. The size of the overlap of the $i_r j_s$ and $i_u j_v$ intervals were also computed, as well as the difference between the minima and the difference between the maxima. Combinations of these quantities, as well as the lone extrema, were tested for their ability to predict the occurrence of dead-ending pairs. Because some of the maxima were very large, the quantities were also compared logarithmically.

Most of the combinations were able to predict dead-ending matrix elements to varying degrees. The best metrics were the fractional interval overlap with respect to each pair, referred to herein as q_{rs} and q_{uv}.

Equation 22

$$q_{rs} = \frac{interval\ overlap}{interval([i_r j_s])} = \frac{\varepsilon_{\max}([i_u j_v]) - \varepsilon_{\min}([i_r j_s])}{\varepsilon_{\max}([i_r j_s]) - \varepsilon_{\min}([i_r j_s])}$$

Equation 23

$$q_{uv} = \frac{interval\ overlap}{interval([i_u j_v])} = \frac{\varepsilon_{\max}([i_u j_v]) - \varepsilon_{\min}([i_r j_s])}{\varepsilon_{\max}([i_u j_v]) - \varepsilon_{\min}([i_u j_v])}$$

These values are calculated using the minima and maxima equations 24, 25, 26 and 27 (see Figure 14):

Equation 24

$$\varepsilon_{\max}([i_r j_s]) = \varepsilon([i_r j_s]) + \sum_{k \neq i \neq j} \max_{l} \varepsilon([i_r j_s], k_l)$$

Equation 25

$$\varepsilon_{\min}([i_r j_s]) = \varepsilon([i_r j_s]) + \sum_{k \neq i \neq j} \min_{t} \varepsilon([i_r j_s], k_l)$$

Equation 26

$$\varepsilon_{\max}([i_u j_v]) = \varepsilon([i_u j_v]) + \sum_{k \neq i \neq j} \max_{t} \varepsilon([i_u j_v], k_l)$$

Equation 27

$$\varepsilon_{\min}([i_u j_v]) = \varepsilon([i_u j_v]) + \sum_{k \neq i \neq j} \min_{t} \varepsilon([i_u j_v], k_l)$$

These metrics were selected because they yield ratios of the occurrence of dead-ending matrix elements to the total occurrence of elements that are higher than any of the other metrics tested. For example, there are very few matrix elements (~2%) for which $q_{rs} > 0.98$, yet these elements produce 30-40% of all of the dead-ending pairs.

Accordingly, the first-order doubles criterion is applied only to those doubles for which $q_{rs} > 0.98$ and $q_{uv} > 0.99$. The sample data analyses predict that by using these two metrics, as many as half of the dead-ending elements may be found by evaluating only two to five percent of the reduced matrix.

Generally, as is more fully described below, single and double DEE, using either or both of original DEE and Goldstein DEE, is run until no further elimination is possible. Usually, convergence is not complete, and further elimination must occur to achieve convergence. This is generally done using "super residue" DEE.

In a preferred embodiment, additional DEE computation is done by the creation of "super residues" or "unification", as is generally described in Desmet, Nature **356**:539-542 (1992); Desmet, *et al.*, The Protein Folding Problem and Tertiary Structure Prediction, Ch. **10**:1-49 (1994); Goldstein, *et al.*, supra. A super residue is a combination of two or more variable residue positions which is then treated as a single residue position. The super residue is then evaluated in singles DEE, and doubles DEE, with either other residue positions or super residues. The disadvantage of super residues is that there are many more rotameric states which must be evaluated; that is, if a first variable residue position has 5 possible rotamers, and a second variable residue position has 4 possible rotamers, there are 20 possible super residue rotamers which must be evaluated. However, these super residues may be eliminated similar to singles, rather than being flagged like pairs.

The selection of which positions to combine into super residues may be done in a variety of ways. In general, random selection of positions for super residues results in inefficient elimination, but it can be done, although this is not preferred. In a preferred embodiment, the first evaluation is the selection of positions for a super residue is the number of rotamers at the position. If the position has too many rotamers, it is never unified into a super residue, as the computation becomes too unwieldy. Thus, only positions with fewer than about 100,000 rotamers are chosen, with less than about 50,000 being preferred and less than about 10,000 being especially preferred.

In a preferred embodiment, the evaluation of whether to form a super residue is done as follows. All possible rotamer pairs are ranked using Equation 28, and the rotamer pair with the highest number is chosen for unification:

Equation 28

$$\log^{(\text{number of super rotamers resulting from the potential unification})}\frac{\text{fraction of flagged pairs}}{}$$

Equation 28 is looking for the pair of positions that has the highest fraction or percentage of flagged pairs but the fewest number of super rotamers. That is, the pair that gives the highest value for Equation 28 is preferably chosen. Thus, if the pair of positions that has the highest number of flagged pairs but also a very large number of super rotamers (that is, the number of rotamers at position i times the number of rotamers at position j), this pair may not be chosen (although it could) over a lower percentage of flagged pairs but fewer super rotamers.

In an alternate preferred embodiment, positions are chosen for super residues that have the highest average energy; that is, for positions i and j, the average energy of all rotamers for i and all rotamers for j is calculated, and the pair with the highest average energy is chosen as a super residue.

Super residues are made one at a time, preferably. After a super residue is chosen, the singles and doubles DEE computations are repeated where the super residue is treated as if it were a regular residue. As for singles and doubles DEE, the elimination of rotamers in the super residue DEE will alter the minimum energy calculations of DEE. Thus, repeating singles and/or doubles DEE can result in further elimination of rotamers.

Figure 3 is a detailed illustration of the processing operations associated with a ranking module 34 of the invention. The calculation and storage of the singles and doubles energies 70 is the first step, although these may be recalculated every time. Step 72 is the optional application of a cutoff, where singles or doubles energies that are too high are eliminated prior to further processing. Either or both of original singles DEE 74 or Goldstein singles DEE 76 may be done, with the elimination of original singles DEE 74 being generally preferred. Once the singles DEE is run, original doubles (78) and/or Goldstein doubles (80) DEE is run. Super residue DEE is then generally run, either original (82) or Goldstein (84) super residue DEE. This preferably results in convergence at a global optimum

sequence. As is depicted in Figure 3, after any step any or all of the previous steps can be rerun, in any order.

The addition of super residue DEE to the computational processing, with repetition of the previous DEE steps, generally results in convergence at the global optimum. Convergence to the global optimum is guaranteed if no cutoff applications are made, although generally a global optimum is achieved even with these steps. In a preferred embodiment, DEE is run until the global optimum sequence is found. That is, the set of optimized protein sequences contains a single member, the global optimum.

In a preferred embodiment, the various DEE steps are run until a managable number of sequences is found, i.e. no further processing is required. These sequences represent a set of optimized protein sequences, and they can be evaluated as is more fully described below. Generally, for computational purposes, a manageable number of sequences depends on the length of the sequence, but generally ranges from about 1 to about 10^{15} possible rotamer sequences.

Alternatively, DEE is run to a point, resulting in a set of optimized sequences (in this context, a set of remainder sequences) and then further compututational processing of a different type may be run. For example, in one embodiment, direct calculation of sequence energy as outlined above is done on the remainder possible sequences. Alternatively, a Monte Carlo search can be run.

In a preferred embodiment, the computation processing need not comprise a DEE computational step. In this embodiment, a Monte Carlo search is undertaken, as is known in the art. See Metropolis *et al.*, J. Chem. Phys. 21:1087 (1953), hereby incorporated by reference. In this embodiment, a random sequence comprising random rotamers is chosen as a start point. In one embodiment, the variable residue positions are classified as core, boundary or surface residues and the set of available residues at each position is thus defined. Then a random sequence is generated, and a random rotamer for each amino acid is chosen. This serves as the starting sequence of the Monte Carlo search. A Monte Carlo search then makes a random jump at one position, either to a different rotamer of the same amino acid or a rotamer of a different amino acid, and then a new sequence energy ($E_{\text{total sequence}}$) is calculated, and if the new sequence energy meets the Boltzmann criteria for acceptance, it is used as the starting point for another jump. If the Boltzmann test fails, another random jump is attempted from the previous sequence. In this way, sequences with lower and lower energies are found, to generate a set of low energy sequences.

If computational processing results in a single global optimum sequence, it is frequently preferred to generate additional sequences in the energy neighborhood of the global solution, which may be ranked. These additional sequences are also optimized protein sequences. The generation of additional optimized sequences is generally preferred so as to evaluate the differences between the theoretical and actual energies of a sequence. Generally, in a preferred embodiment, the set of sequences is at least about 75% homologous to each other, with at least about 80% homologous being preferred, at least about 85% homologous being particularly preferred, and at least about 90%

being especially preferred. In some cases, homology as high as 95% to 98% is desirable. Homology in this context means sequence similarity or identity, with identity being preferred. Identical in this context means identical amino acids at corresponding positions in the two sequences which are being compared. Homology in this context includes amino acids which are identical and those which are similar (functionally equivalent). This homology will be determined using standard techniques known in the art, such as the Best Fit sequence program described by Devereux, *et al.*, Nucl. Acid Res., **12**:387-395 (1984), or the BLASTX program (Altschul, *et al.*, J. Mol. Biol., **215**:403-410 (1990)) preferably using the default settings for either. The alignment may include the introduction of gaps in the sequences to be aligned. In addition, for sequences which contain either more or fewer amino acids than an optimum sequence, it is understood that the percentage of homology will be determined based on the number of homologous amino acids in relation to the total number of amino acids. Thus, for example, homology of sequences shorter than an optimum will be determined using the number of amino acids in the shorter sequence.

Once optimized protein sequences are identified, the processing of Figure 2 optionally proceeds to step 56 which entails searching the protein sequences. This processing may be implemented with the search module 36. The search module 36 is a set of computer code that executes a search strategy. For example, the search module 36 may be written to execute a Monte Carlo search as described above. Starting with the global solution, random positions are changed to other rotamers allowed at the particular position, both rotamers from the same amino acid and rotamers from different amino acids. A new sequence energy ($E_{total\ sequence}$) is calculated, and if the new sequence energy meets the Boltzmann criteria for acceptance, it is used as the starting point for another jump. See Metropolis *et al.*, 1953, supra, hereby incorporated by reference. If the Boltzmann test fails, another random jump is attempted from the previous sequence. A list of the sequences and their energies is maintained during the search. After a predetermined number of jumps, the best scoring sequences may be output as a rank-ordered list. Preferably, at least about 10^6 jumps are made, with at least about 10^7 jumps being preferred and at least about 10^8 jumps being particularly preferred. Preferably, at least about 100 to 1000 sequences are saved, with at least about 10,000 sequences being preferred and at least about 100,000 to 1,000,000 sequences being especially preferred. During the search, the temperature is preferably set to 1000 K.

Once the Monte Carlo search is over, all of the saved sequences are quenched by changing the temperature to 0 K, and fixing the amino acid identity at each position. Preferably, every possible rotamer jump for that particular amino acid at every position is then tried.

The computational processing results in a set of optimized protein sequences. These optimized protein sequences are generally, but not always, significantly different from the wild-type sequence from which the backbone was taken. That is, each optimized protein sequence preferably comprises at least about 5-10% variant amino acids from the starting or wild-type sequence, with at least about 15-20% changes being preferred and at least about 30% changes being particularly preferred.

These sequences can be used in a number of ways. In a preferred embodiment, one, some or all of the optimized protein sequences are constructed into designed proteins, as show with step 58 of Figure 2. Thereafter, the protein sequences can be tested, as shown with step 60 of the Figure 2. Generally, this can be done in one of two ways.

In a preferred embodiment, the designed proteins are chemically synthesized as is known in the art. This is particularly useful when the designed proteins are short, preferably less than 150 amino acids in length, with less than 100 amino acids being preferred, and less than 50 amino acids being particularly preferred, although as is known in the art, longer proteins can be made chemically or enzymatically.

In a preferred embodiment, particularly for longer proteins or proteins for which large samples are desired, the optimized sequence is used to create a nucleic acid such as DNA which encodes the optimized sequence and which can then be cloned into a host cell and expressed. Thus, nucleic acids, and particularly DNA, can be made which encodes each optimized protein sequence. This is done using well known procedures. The choice of codons, suitable expression vectors and suitable host cells will vary depending on a number of factors, and can be easily optimized as needed.

Once made, the designed proteins are experimentally evaluated and tested for structure, function and stability, as required. This will be done as is known in the art, and will depend in part on the original protein from which the protein backbone structure was taken. Preferably, the designed proteins are more stable than the known protein that was used as the starting point, although in some cases, if some constaints are placed on the methods, the designed protein may be less stable. Thus, for example, it is possible to fix certain residues for altered biological activity and find the most stable sequence, but it may still be less stable than the wild type protein. Stable in this context means that the new protein retains either biological activity or conformation past the point at which the parent molecule did. Stability includes, but is not limited to, thermal stability, i.e. an increase in the temperature at which reversible or irreversible denaturing starts to occur; proteolytic stability, i.e. a decrease in the amount of protein which is irreversibly cleaved in the presence of a particular protease (including autolysis); stability to alterations in pH or oxidative conditions; chelator stability; stability to metal ions; stability to solvents such as organic solvents, surfactants, formulation chemicals; etc.

In a preferred embodiment, the modelled proteins are at least about 5% more stable than the original protein, with at least about 10% being preferred and at least about 20-50% being especially preferred.

The results of the testing operations may be computationally assessed, as shown with step 62 of Figure 2. An assessment module 38 may be used in this operation. That is, computer code may be prepared to analyze the test data with respect to any number of metrices.

At this processing juncture, if the protein is selected (the yes branch at block 64) then the protein is utilized (step 66), as discussed below. If a protein is not selected, the accumulated information may

be used to alter the ranking module 34, and/or step 56 is repeated and more sequences are searched.

In a preferred embodiment, the experimental results are used for design feedback and design optimization.

Once made, the proteins of the invention find use in a wide variety of applications, as will be appreciated by those in the art, ranging from industrial to pharmocological uses, depending on the protein. Thus, for example, proteins and enzymes exhibiting increased thermal stability may be used in industrial processes that are frequently run at elevated temperatures, for example carbohydrate processing (including saccharification and liquifaction of starch to produce high fructose corn syrup and other sweetners), protein processing (for example the use of proteases in laundry detergents, food processing, feed stock processing, baking, etc.), etc. Similarly, the methods of the present invention allow the generation of useful pharmaceutical proteins, such as analogs of known proteinaceous drugs which are more thermostable, less proteolytically sensitive, or contain other desirable changes.

The following examples serve to more fully describe the manner of using the above-described invention, as well as to set forth the best modes contemplated for carrying out various aspects of the invention. It is understood that these examples in no way serve to limit the true scope of this invention, but rather are presented for illustrative purposes. All references cited herein are explicitly incorporated by reference in their entirety.

EXAMPLES

Example 1

Protein Design Using van der Waals and Atomic Solvation Scoring Functions with DEE

A cyclical design strategy was developed that couples theory, computation and experimental testing in order to address the problems of specificity and learning (Figure 4). Our protein design automation (PDA) cycle is comprised of four components: a design paradigm, a simulation module, experimental testing and data analysis. The design paradigm is based on the concept of inverse folding (Pabo, Nature **301**:200 (1983); Bowie, *et al.*, Science **253**:164-170 (1991)) and consists of the use of a fixed backbone onto which a sequence of side-chain rotamers can be placed, where rotamers are the allowed conformations of amino acid side chains (Ponder, *et al.*, (1987) (supra)). Specific tertiary interactions based on the three dimensional juxtaposition of atoms are used to determine the sequences that will potentially best adopt the target fold. Given a backbone geometry and the possible rotamers allowed for each residue position as input, the simulation must generate as output a rank ordered list of solutions based on a cost function that explicitly considers the atom positions in the various rotamers. The principle obstacle is that a fixed backbone comprised of n residues and m possible rotamers per residue (all rotamers of all allowed amino acids) results in m^n possible arrangements of the system, an immense number for even small design problems. For example, to

consider 50 rotamers at 15 positions results in over 10^{25} sequences, which at an evaluation rate of 10^9 sequences per second (far beyond current capabilities) would take 10^9 *years* to exhaustively search for the global minimum. The synthesis and characterization of a subset of amino acid sequences presented by the simulation module generates experimental data for the analysis module. The analysis section discovers correlations between calculable properties of the simulated structures and the experimental observables. The goal of the analysis is to suggest *quantitative* modifications to the simulation and in some cases to the guiding design paradigm. In other words, the cost function used in the simulation module describes a theoretical potential energy surface whose horizontal axis comprises all possible solutions to the problem at hand. This potential energy surface is not guaranteed to match the actual potential energy surface which is determined from the experimental data. In this light, the goal of the analysis becomes the correction of the simulation cost function in order to create better agreement between the theoretical and actual potential energy surfaces. If such corrections can be found, then the output of subsequent simulations will be amino acid sequences that better achieve the target properties. This design cycle is generally applicable to any protein system and, by removing the subjective human component, allows a largely unbiased approach to protein design, i.e. protein design automation.

The PDA side-chain selection algorithm requires as input a backbone structure defining the desired fold. The task of designing a sequence that takes this fold can be viewed as finding an optimal arrangement of amino acid side chains relative to the given backbone. It is not sufficient to consider *only* the identity of an amino acid when evaluating sequences. In order to correctly account for the geometric specificity of side-chain placement, all possible conformations of each side chain must also be examined. Statistical surveys of the protein structure database (Ponder, *et al.*, supra) have defined a discrete set of allowed conformations, called rotamers, for each amino acid side chain. We use a rotamer library based on the Ponder and Richards library to define allowed conformations for the side chains in PDA.

Using a rotamer description of side chains, an optimal sequence for a backbone can be found by screening all possible sequences of rotamers, where each backbone position can be occupied by each amino acid in all its possible rotameric states. The discrete nature of rotamer sets allows a simple calculation of the number of rotamer sequences to be tested. A backbone of length n with m possible rotamers per position will have m^n possible rotamer sequences. The size of the search space grows exponentially with sequence length which for typical values of n and m render intractable an exhaustive search. This combinatorial "explosion" is the primary obstacle to be overcome in the simulation phase of PDA.

Simulation algorithm: An extension of the Dead End Elimination (DEE) theorem was developed (Desmet, *et al.*, (1992((supra); Desmet, *et al.*, (1994) (supra); Goldstein, (1994) (supra) to solve the combinatorial search problem. The DEE theorem is the basis for a very fast discrete search algorithm that was designed to pack protein side chains on a fixed backbone with a known sequence. Side chains are described by rotamers and an atomistic forcefield is used to score rotamer arrangements.

The DEE theorem guarantees that if the algorithm converges, the *global* optimum packing is found. The DEE method is readily extended to our inverse folding design paradigm by releasing the constraint that a position is limited to the rotamers of a single amino acid. This extension of DEE greatly increases the number of rotamers at each position and requires a significantly modified implementation to ensure convergence, described more fully herein. The guarantee that only the global optimum will be found is still valid, and in our extension means that the globally optimal sequence is found in its optimal conformation.

DEE was implemented with a novel addition to the improvements suggested by Goldstein (Goldstein, (1994) (supra)). As has been noted, exhaustive application of the R=1 rotamer elimination and R=0 rotamer-pair flagging equations and limited application of the R=1 rotamer-pair flagging equation routinely fails to find the global solution. This problem can be overcome by unifying residues into "super residues" (Desmet, *et al.*, (1992((supra); Desmet, *et al.*, (1994) (supra); Goldstein, (1994) (supra). However, unification can cause an unmanageable increase in the number of super rotamers per super residue position and can lead to intractably slow performance since the computation time for applying the R=1 rotamer-pair flagging equation increases as the fourth power of the number of rotamers. These problems are of particular importance for protein design applications given the requirement for large numbers of rotamers per residue position. In order to limit memory size and to increase performance, we developed a heuristic that governs which residues (or super residues) get unified and the number of rotamer (or super rotamer) pairs that are included in the R=1 rotamer-pair flagging equation. A program called PDA_DEE was written that takes a list of rotamer energies from PDA_SETUP and outputs the global minimum sequence in its optimal conformation with its energy.

Scoring functions: The rotamer library used was similar to that used by Desmet and coworkers (Desmet, *et al.*, (1992) (supra)). χ_1 and χ_2 angle values of rotamers for all amino acids except Met, Arg and Lys were expanded plus and minus one standard deviation about the mean value from the Ponder and Richards library (supra) in order to minimize possible errors that might arise from the discreteness of the library. c_3 and c_4 angles that were undetermined from the database statistics were assigned values of 0° and 180° for Gln and 60°, -60° and 180° for Met, Lys and Arg. The number of rotamers per amino acid is: Gly, 1; Ala, 1; Val, 9; Ser, 9; Cys, 9; Thr, 9; Leu, 36; Ile, 45; Phe, 36; Tyr, 36; Trp, 54; His, 54; Asp, 27; Asn, 54; Glu, 69; Gln, 90; Met, 21; Lys, 57; Arg, 55. The cyclic amino acid Pro was not included in the library. Further, all rotamers in the library contained explicit hydrogen atoms. Rotamers were built with bond lengths and angles from the Dreiding forcefield (Mayo, *et al.*, J. Phys. Chem. 94:8897 (1990)).

The initial scoring function for sequence arrangements used in the search was an atomic van der Waals potential. The van der Waals potential reflects excluded volume and steric packing interactions which are important determinants of the specific three dimensional arrangement of protein side chains. A Lennard-Jones 12-6 potential with radii and well depth parameters from the Dreiding forcefield was used for van der Waals interactions. Non-bonded interactions for atoms connected by one or two bonds were not considered. van der Waals radii for atoms connected by

three bonds were scaled by 0.5. Rotamer/rotamer pair energies and rotamer/template energies were calculated in a manner consistent with the published DEE algorithm (Desmet, *et al.*, (1992) (supra)). The template consisted of the protein backbone and the side chains of residue positions not to be optimized. No intra-side-chain potentials were calculated. This scheme scored the packing geometry and eliminated bias from rotamer internal energies. Prior to DEE, all rotamers with template interaction energies greater than 25 kcal/mol were eliminated. Also, any rotamer whose interaction was greater than 25 kcal/mol with all other rotamers at another residue position was eliminated. A program called PDA_SETUP was written that takes as input backbone coordinates, including side chains for positions not optimized, a rotamer library, a list of positions to be optimized and a list of the amino acids to be considered at each position. PDA_SETUP outputs a list of rotamer/template and rotamer/rotamer energies.

The pairwise solvation potential was implemented in two components to remain consistent with the DEE methodology: rotamer/template and rotamer/rotamer burial. For the rotamer/template buried area, the reference state was defined as the rotamer in question at residue *i* with the backbone atoms only of residues *i*-1, *i* and *i*+1. The area of the side chain was calculated with the backbone atoms excluding solvent but not counted in the area. The folded state was defined as the area of the rotamer in question at residue *i*, but now in the context of the entire template structure including non-optimized side chains. The rotamer/template buried area is the difference between the reference and the folded states. The rotamer/rotamer reference area is simply the sum of the areas of the isolated rotamers. The folded state is the area of the two rotamers placed in their relative positions on the protein scaffold but with no template atoms present. The Richards definition of solvent accessible surface area (Lee & Richards, 1971, supra) was used, with a probe radius of 1.4 Å and Drieding van der Waals radii. Carbon and sulfur, and all attached hydrogens, were considered nonpolar. Nitrogen and oxygen, and all attached hydrogens, were considered polar. Surface areas were calculated with the Connolly algorithm using a dot density of 10 $Å^{-2}$ (Connolly, (1983) (supra)). In more recent implementations of PDA_SETUP, the MSEED algorithm of Scheraga has been used in conjunction with the Connolly algorithm to speed up the calculation (Perrot, *et al.*, J. Comput. Chem. 13:1-11 (1992)).

Monte Carlo search: Following DEE optimization, a rank ordered list of sequences was generated by a Monte Carlo search in the neighborhood of the DEE solution. This list of sequences was necessary because of possible differences between the theoretical and actual potential surfaces. The Monte Carlo search starts at the global minimum sequence found by DEE. A residue was picked randomly and changed to a random rotamer selected from those allowed at that site. A new sequence energy was calculated and, if it met the Boltzman criteria for acceptance, the new sequence was used as the starting point for another jump. If the Boltzman test failed, then another random jump was attempted from the previous sequence. A list of the best sequences found and their energies was maintained throughout the search. Typically 10^6 jumps were made, 100 sequences saved and the temperature was set to 1000 K. After the search was over, all of the saved sequences were quenched by changing the temperature to 0 K, fixing the amino acid identity and

trying every possible rotamer jump at every position. The search was implemented in a program called PDA_MONTE whose input was a global optimum solution from PDA_DEE and a list of rotamer energies from PDA_SETUP. The output was a list of the best sequences rank ordered by their score. PDA_SETUP, PDA_DEE and PDA_MONTE were implemented in the CERIUS2 software development environment (Biosym/Molecular Simulations, San Diego, CA).

PDA_SETUP, PDA_DEE, and PDA_MONTE were implemented in the CERIUS2 software development environment (Biosym/Molecular Simulations, San Diego, CA).

Model system and experimental testing: The homodimeric coiled coil of α helices was selected as the initial design target. Coiled coils are readily synthesized by solid phase techniques and their helical secondary structure and dimeric tertiary organization ease characterization. Their sequences display a seven residue periodic HP pattern called a heptad repeat, (**a b c d e f g**) (Cohen & Parry, Proteins Struc. Func. Genet. 7:1-15 (1990)). The **a** and **d** positions are usually hydrophobic and buried at the dimer interface while the other positions are usually polar and solvent exposed (Figure 5). The backbone needed for input to the simulation module was taken from the crystal structure of GCN4-p1 (O'Shea, *et al.*, Science 254:539 (1991)). The 16 hydrophobic **a** and **d** positions were optimized in the crystallographically determined fixed field of the rest of the protein. Homodimer sequence symmetry was enforced, only rotamers from hydrophobic amino acids (A, V, L, I, M, F, Y and W) were considered and the asparagine at an **a** position, Asn 16, was not optimized.

Homodimeric coiled coils were modeled on the backbone coordinates of GCN4-p1, PDB ascension code 2ZTA (Bernstein, *et al.*, J. Mol. Biol. 112:535 (1977); O'Shea, *et al.*, supra). Atoms of all side chains not optimized were left in their crystallographically determined positions. The program BIOGRAF (Biosym/Molecular Simulations, San Diego, CA) was used to generate explicit hydrogens on the structure which was then conjugate gradient minimized for 50 steps using the Dreiding forcefield. The HP pattern was enforced by only allowing hydrophobic amino acids into the rotamer groups for the optimized **a** and **d** positions. The hydrophobic group consisted of Ala, Val, Leu, Ile, Met, Phe, Tyr and Trp for a total of 238 rotamers per position. Homodimer symmetry was enforced by penalizing by 100 kcal/mol rotamer pairs that violate sequence symmetry. Different rotamers of the same amino acid were allowed at symmetry related positions. The asparagine that occupies the **a** position at residue 16 was left in the template and not optimized. A 10^6 step Monte Carlo search run at a temperature of 1000 K generated the list of candidate sequences rank ordered by their score. To test reproducibility, the search was repeated three times with different random number seeds and all trials provided essentially identical results. The Monte Carlo searches took about 90 minutes. All calculations in this work were performed on a Silicon Graphics 200 MHz R4400 processor.

Optimizing the 16 **a** and **d** positions each with 238 possible hydrophobic rotamers results in 238^{16} or 10^{38} rotamer sequences. The DEE algorithm finds the global optimum in three minutes, including rotamer energy calculation time. The DEE solution matches the naturally occurring GCN4-p1 sequence of **a** and **d** residues for all of the 16 positions. A one million step Monte Carlo search run at a temperature of 1000 K generated the list of sequences rank ordered by their score. To test

reproducibility, the search was repeated three times with different random number seeds and all trials provided essentially identical results. The second best sequence is a Val 30 to Ala mutation and lies three kcal/mol above the ground state sequence. Within the top 15 sequences up to six mutations from the ground state sequence are tolerated, indicating that a variety of packing arrangements are available even for a small coiled coil. Eight sequences with a range of stabilities were selected for experimental testing, including six from the top 15 and two more about 15 kcal/mol higher in energy, the 56th and 70th in the list (Table 1).

TABLE 1

Name	Sequence	Rank	Energy
PDA-3H[b]	RMKQLEDKVEELLSKNYHLENEVARLKKLVGER (SEQ ID NO:23)	1	-118.1
PDA-3A	RMKQLEDKVEELLSKNYHLENEVARLKKLAGER (SEQ ID NO:24)	2	-115.3
PDA-3G	RMKQLEDKVEELLSKNYHLENEMARLKKLVGER (SEQ ID NO:25)	5	-112.8
PDA-3B	RLKQMEDKVEELLSKNYHLENEVARLKKLVGER (SEQ ID NO:26)	6	-112.6
PDA-3D	RLKQMEDKVEELLSKNYHLENEVARLKKLAGER (SEQ ID NO:27)	13	-109.7
PDA-3C	RMKQWEDKAEELLSKNYHLENEVARLKKLVGER (SEQ ID NO:28)	14	-109.6
PDA-3F	RMKQFEDKVEELLSKNYHLENEVARLKKLVGER (SEQ ID NO:29)	56	-103.9
PDA-3E	RMKQLEDKVEELLSKNYHAENEVARLKKLVGER (SEQ ID NO:30)	70	-103.1

Thirty-three residue peptides were synthesized on an Applied Biosystems Model 433A peptide synthesizer using Fmoc chemistry, HBTU activation and a modified Rink amide resin from Novabiochem. Standard 0.1 mmol coupling cycles were used and amino termini were acetylated. Peptides were cleaved from the resin by treating approximately 200 mg of resin with 2 mL trifluoroacetic acid (TFA) and 100 µL water, 100 µL thioanisole, 50 µL ethanedithiol and 150 mg phenol as scavengers. The peptides were isolated and purified by precipitation and repeated washing with cold methyl tert-butyl ether followed by reverse phase HPLC on a Vydac C8 column (25 cm by 22 mm) with a linear acetonitrile-water gradient containing 0.1% TFA. Peptides were then lyophilized and stored at -20 °C until use. Plasma desorption mass spectrometry found all molecular weights to be within one unit of the expected masses.

Circular dichroism CD spectra were measured on an Aviv 62DS spectrometer at pH 7.0 in 50 mM phosphate, 150 mM NaCl and 40 µM peptide. A 1 mm pathlength cell was used and the temperature was controlled by a thermoelectric unit. Thermal melts were performed in the same buffer using two degree temperature increments with an averaging time of 10 s and an equilibration time of 90 s. T_m values were derived from the ellipticity at 222 nm ($[\theta]_{222}$) by evaluating the minimum of the $d[\theta]_{222}/dT^{-1}$ versus T plot (Cantor & Schimmel, Biophysical Chemistry. New York: W. H. Freemant and Company, 1980). The T_m's were reproducible to within one degree. Peptide concentrations were determined from the tyrosine absorbance at 275 nm (Huyghues-Despointes, *et al.*, supra).

Size exclusion chromatography: Size exclusion chromatography was performed with a Synchropak GPC 100 column (25 cm by 4.6 mm) at pH 7.0 in 50 mM phosphate and 150 mM NaCl at 0 °C. GCN4-p1 and p-LI (Harbury, *et al.*, Science 262:1401 (1993)) were used as size standards. 10 µl

injections of 1 mM peptide solution were chromatographed at 0.20 ml/min and monitored at 275 nm. Peptide concentrations were approximately 60 μM as estimated from peak heights. Samples were run in triplicate.

The designed **a** and **d** sequences were synthesized as above using the GCN4-p1 sequence for the **b c** and **e f g** positions. Standard solid phase techniques were used and following HPLC purification, the identities of the peptides were confirmed by mass spectrometry. Circular dichroism spectroscopy (CD) was used to assay the secondary structure and thermal stability of the designed peptides. The CD spectra of all the peptides at 1 °C and a concentration of 40 mM exhibit minima at 208 and 222 nm and a maximum at 195 nm, which are diagnostic for α helices (data not shown). The ellipticity values at 222 nm indicate that all of the peptides are >85% helical (approximately -28000 deg cm^2/dmol), with the exception of PDA-3C which is 75% helical at 40 mM but increases to 90% helical at 170 mM (Table 2).

Table 2. CD data and calculated structural properties of the PDA peptides.

Name	$-[\theta]_{222}$ (deg cm^2/dmol)	T_m (°C)	E_{MC} (kcal/ mol)	ΔA_{np} (Å^2)	ΔA_p (Å^2)	Vol (Å^3)	Rot bonds	E_{CQ} (kcal/ mol)	E_{CG} (kcal/ mol)	E_{vdW} (kcal/ mol)	Npb	Pb
PDA-3 H	33000	57	-118.1	2967	2341	1830	28	-234	-308	409	207	128
PDA-3 A	30300	48	-115.3	2910	2361	1725	26	-232	-312	400	203	128
PDA-3 B	28200	47	-112.6	2977	2372	1830	28	-242	-306	379	210	127
PDA-3 G	30700	47	-112.8	3003	2383	1878	32	-240	-309	439	212	128
PDA-3 F	28800	39	-103.9	3000	2336	1872	28	-188	-302	420	212	128
PDA-3 D	27800	39	-109.7	2920	2392	1725	26	-240	-310	370	206	127
PDA-3 C	24100	26	-109.6	2878	2400	1843	26	-149	-304	398	215	129
PDA-3 E	27500	24	-103.1	2882	2361	1674	24	-179	-309	411	203	127

*E_{MC} is the Monte Carlo energy; ΔA_{np} and ΔA_p are the changes in solvent accessible non-polar and polar surface areas upon folding, respectively; E_{CQ} is the electrostatic energy using equilibrated charges; E_{CG} is the electrostatic energy using Gasteiger charges; E_{vdW} is the van der Waals energy; Vol is the side chain van der Waals volume; Rot bonds is the number of side chain rotatable bonds

(excluding methyl rotors); Npb and Pb are the number of buried non-polar and polar atoms, respectively.

The melting temperatures (T_m's) show a broad range of values (data not shown), with 6 of the 8 peptides melting at greater than physiological temperature. Also, the T_m's were not correlated to the number of sequence differences from GCN4-p1. Single amino acid changes resulted in some of the most and least stable peptides, demonstrating the importance of specificity in sequence selection.

Size exclusion chromatography confirmed the dimeric nature of these designed peptides. Using coiled coil peptides of known oligomerization state as standards, the PDA peptides migrated as dimers. This result is consistent with the appearance of β-branched residues at **a** positions and leucines at **d** positions, which have been shown previously to favor dimerization over other possible oligomerization states (Harbury, *et al.*, supra).

The characterization of the PDA peptides demonstrates the successful design of several stable dimeric helical coiled coils. The sequences were automatically generated in the context of the design paradigm by the simulation module using well-defined inputs that explicitly consider the HP patterning and steric specificity of protein structure. Two dimensional nuclear magnetic resonance experiments aimed at probing the specificity of the tertiary packing are the focus of further studies on these peptides. Initial experiments show significant protection of amide protons from chemical exchange and chemical shift dispersion comparable to GCN4-p1 (unpublished results) (Oas, *et al.*, Biochemistry 29:2891 (1990)); Goodman & Kim, Biochem. 30:11615 (1991)).

Data analysis and design feedback A detailed analysis of the correspondence between the theoretical and experimental potential surfaces, and hence an estimate of the accuracy of the simulation cost function, was enabled by the collection of experimental data. Using thermal stability as a measure of design performance, melting temperatures of the PDA peptides were plotted against the sequence scores found in the Monte Carlo search (Figure 6). The modest correlation, 0.67, in the plot shows that while an exclusively van der Waals scoring function can screen for stable sequences, it does not accurately predict relative stabilities. In order to address this issue, correlations between calculated structural properties and T_m's were systematically examined using quantitative structure activity relationships (QSAR), which is a statistical technique commonly used in structure based drug design (Hopfinger, J. Med. Chem. 28:1133 (1985)).

Table 2 lists various molecular properties of the PDA peptides in addition to the van der Waals based Monte Carlo scores and the experimentally determined T_m's. A wide range of properties was examined, including molecular mechanics components, such as electrostatic energies, and geometric measures, such as volume. The goal of QSAR is the generation of equations that closely approximate the experimental quantity, in this case T_m, as a function of the calculated properties. Such equations suggest which properties can be used in an improved cost function. The PDA analysis module employs genetic function approximation (GFA) (Rogers & Hopfinger, J. Chem. Inf. Comput. Scie. 34:854 (1994)), a novel method to optimize QSAR equations that selects which

properties are to be included and the relative weightings of the properties using a genetic algorithm. GFA accomplishes an efficient search of the space of possible equations and robustly generates a list of equations ranked by their correlation to the data.

Equations are scored by lack of fit (LOF), a weighted least square error measure that resists overfitting by penalizing equations with more terms (Rogers & Hopfinger, supra). GFA optimizes both the length and the composition of the equations and, by generating a set of QSAR equations, clarifies combinations of properties that fit well and properties that recur in many equations. All of the top five equations that correct the simulation energy (E_{MC}) contain burial of nonpolar surface area, ΔA_{np} (Table 3).

Table 3. Top five QSAR equations generated by GFA with LOF, correlation coefficient and cross validation scores.

QSAR equation	LOF	r^2	CV r^2
$-1.44*E_{MC} + 0.14*\Delta A_{np} - 0.73*Npb$	16.23	.98	.78
$-1.78*E_{MC} + 0.20*\Delta A_{np} - 2.43*Rot$	23.13	.97	.75
$-1.59*E_{MC} + 0.17*\Delta A_{np} - 0.05*Vol$	24.57	.97	.36
$-1.54*E_{MC} + 0.11*\Delta A_{np}$	25.45	.91	.80
$-1.60*E_{MC} + 0.09*\Delta A_{np} - 0.12*\Delta A_{p}$	33.88	.96	.90

ΔA_{np} and ΔA_{p} are nonpolar and polar surface buried upon folding, respectively. Vol is side chain volume, Npb is the number of buried nonpolar atoms and Rot is the number of buried rotatable bonds.

The presence of ΔA_{np} in all of the top equations, in addition to the low LOF of the QSAR containing only E_{MC} and ΔA_{np}, strongly implicates nonpolar surface burial as a critical property for predicting peptide stability. This conclusion is not surprising given the role of the hydrophobic effect in protein energetics (Dill, Biochem. 29:7133 (1990)).

Properties were calculated using BIOGRAF and the Dreiding forcefield. Solvent accessible surface areas were calculated with the Connolly algorithm (Connolly, (1983) (supra)) using a probe radius of 1.4 Å and a dot density of 10 $Å^{-2}$. Volumes were calculated as the sum of the van der Waals volumes of the side chains that were optimized. The number of buried polar and nonpolar heavy atoms were defined as atoms, with their attached hydrogens, that expose less than 5 $Å^2$ in the surface area calculation. Electrostatic energies were calculated using a dielectric of one and no cutoff was set for calculation of non-bonded energies. Charge equilibration charges (Rappe & Goddard III, J. Phys. Chem. 95:3358 (1991) and Gasteiger (Gasteiger & Marsili, Tetrahedron 36:3219 (1980) charges were used to generate electrostatic energies. Charge equilibration charges were manually adjusted to

provide neutral backbones and neutral side chains in order to prevent spurious monopole effects. The selection of properties was limited by the requirement that properties could not be highly correlated. Correlated properties cannot be differentiated by QSAR techniques and only create redundancy in the derived relations.

Genetic function approximation (GFA) was performed in the CERIUS2 simulation package version 1.6 (Biosym/Molecular Simulations, San Diego, CA). An initial population of 300 equations was generated consisting of random combinations of three properties. Only linear terms were used and initial coefficients were determined by least squares regression for each set of properties. Redundant equations were eliminated and 10000 generations of random crossover mutations were performed. If a child had a better score than the worst equation in the population, the child replaced the worst equation. Also, mutation operators that add or remove terms had a 50% probability of being applied each generation, but these mutations were only accepted if the score was improved. No equation with greater than three terms was allowed. Equations were scored during evolution using the lack of fit (LOF) parameter, a scaled least square error (LSE) measure that penalizes equations with more terms and hence resists overfitting. LOF is defined as:

$$LOF = \frac{LSE}{(1 - \frac{2C}{M})^2}$$

where c is the number of terms in the equation and M is the number of data points. Five different randomized runs were done and the final equation populations were pooled. Only equations containing the simulation energy, E_{MC}, were considered which resulted in 108 equations ranked by their LOF.

To assess the predictive power of these QSAR equations, as well as their robustness, cross validation analysis was carried out. Each peptide was sequentially removed from the data set and the coefficients of the equation in question were refit. This new equation was then used to predict the withheld data point. When all of the data points had been predicted in this manner, their correlation to the measured T_m's was computed (Table 3). Only the $E_{MC}/\Delta A_{np}$ QSAR and the $E_{MC}/\Delta A_{np}/\Delta A_p$ QSAR performed well in cross validation. The $E_{MC}/\Delta A_{np}$ equation could not be expected to fit the data as smoothly as QSAR's with three terms and hence had a lower cross validated r^2. However, all other two term QSAR's had LOF scores greater than 48 and cross validation correlations less than 0.55 (data not shown). The QSAR analysis independently predicted with no subjective bias that consideration of nonpolar and polar surface area burial is necessary to improve the simulation. This result is consistent with previous studies on atomic solvation potentials (Eisenberg, *et al.*, (1986) (supra); Wesson, *et al.*, Protein Sci. 1:227 (1992)). Further, simpler structural measures, such as number of buried atoms, that reflect underlying principles such as hydrophobic solvation (Chan, *et al.*, Science 267:1463 (1995)) were not deemed as significant by the QSAR analysis. These results

justify the cost of calculating actual surface areas, though in some studies simpler potentials have been shown to perform well (van Gunsteren, *et al.* J. Mol. Biol 227:389 (1992)).

Δa_{np} and ΔA_p were introduced into the simulation module to correct the cost function. Contributions to surface burial from rotamer/template and rotamer/rotamer contacts were calculated and used in the interaction potential. Independently counting buried surface from different rotamer pairs, which is necessary in DEE, leads to overestimation of burial because the radii of solvent accessible surfaces are much larger than the van der Waals contact radii and hence can overlap greatly in a close packed protein core. To account for this discrepancy, the areas used in the QSAR were recalculated using the pairwise area method and a new $E_{MC}/\Delta A_{np}/\Delta A_p$ QSAR equation was generated. The ratios of the E_{MC} coefficient to the ΔA_{np} and ΔA_p coefficients are scale factors that are used in the simulation module to convert buried surface area into energy, i.e. atomic solvation parameters. Thermal stabilities are predicted well by this cost function (Figure 6B). In addition, the improved cost function still predicts the naturally occurring GCN4-p1 sequence as the ground state. The surface area to energy scale factors, 16 cal/mol/Å^2 favoring nonpolar area burial and 86 cal/mol/Å^2 opposing polar area burial, are similar in sign, scale and relative magnitude to solvation potential parameters derived from small molecule transfer data (Wesson & Eisenberg, supra).

λ repressor mutants: To demonstrate the generality of the cost function, other proteins were examined using the simulation module. A library of core mutants of the DNA binding protein λ repressor has been extensively characterized by Sauer and coworkers (Lim & Sauer, J. Mol. Biol. 219:359 (1991)). Template coordinates were taken from PDB file 1LMB (Beamer & Pabo, J. Mol. Biol. 227:177 (1992)). The subunit designated chain 4 in the PDB file was removed from the context of the rest of the structure (accompanying subunit and DNA) and using BIOGRAF explicit hydrogens were added. The hydrophobic residues with side chains within 5 Å of the three mutation sites (V36 M40 V47) are Y22, L31, A37, M42, L50, F51, L64, L65, I68 and L69. All of these residues are greater than 80% buried except for M42 which is 65% buried and L64 which is 45% buried. A37 only has one possible rotamer and hence was not optimized. The other nine residues in the 5 Å sphere were allowed to take any rotamer conformation of their amino acid ("floated"). The mutation sites were allowed any rotamer of the amino acid sequence in question. Depending on the mutant sequence, 5 x 10^{16} to 7 x 10^{18} conformations were possible. Rotamer energy and DEE calculation times were 2 to 4 minutes. The combined activity score is that of Hellinga and Richards (Hellinga, *et al.*, (1994) (supra)). Seventy-eight of the 125 possible combinations were generated. Also, this dataset has been used to test several computational schemes and can serve as a basis for comparing different forcefields (Lee & Levitt, Nature 352:448 (1991); van Gunsteren & Mark, supra; Hellinga, *et al.*, (1994) (supra)). The simulation module, using the cost function found by QSAR, was used to find the optimal conformation and energy for each mutant sequence. All hydrophobic residues within 5 Å of the three mutation sites were also left free to be relaxed by the algorithm. This 5 Å sphere contained 12 residues, a significantly larger problem than previous efforts (Lee & Levitt, supra; Hellinga, (1994) (supra)), that were rapidly optimized by the DEE component of the simulation module. The rank correlation of the predicted energy to the combined activity score proposed by Hellinga and Richards

is shown in Figure 7. The wildtype has the lowest energy of the 125 possible sequences and the correlation is essentially equivalent to previously published results which demonstrates that the QSAR corrected cost function is not specific for coiled coils and can model other proteins adequately.

Example 2

Automated design of the surface positions of protein helices

GCN4-pl, a homodimeric coiled coil, was again selected as the model system because it can be readily synthesized by solid phase techniques and its helical secondary structure and dimeric tertiary organization ease characterization. The sequences of homodimeric coiled coils display a seven residue periodic hydrophobic and polar pattern called a heptad repeat, (**a b c d e f g**) (Cohen & Parry, supra). The **a** and **d** positions are buried at the dimer interface and are usually hydrophobic, whereas the **b**, **c**, **e**, **f**, and **g** positions are solvent exposed and usually polar (Figure 5). Examination of the crystal structure of GCN4-p1 (O'Shea, *et al.*, supra) shows that the **b**, **c**, and **f** side chains extend into solvent and expose at least 55% of their surface area. In contrast, the **e** and **g** residues bury from 50 to 90% of their surface area by packing against the **a** and **d** residues of the opposing helix. We selected the 12 **b**, **c**, and **f** residue positions for surface sequence design: positions 3, 4, 7, 10, 11, 14, 17, 18, 21, 24, 25, and 28 using the numbering from PDB entry 2zta (Bernstein, *et al.*, J. Mol. Biol. 112:535 (1977)). The remainder of the protein structure, including all other side chains and the backbone, was used as the template for sequence selection calculations. The symmetry of the dimer and lack of interactions of surface residues between the subunits allowed independent design of each subunit, thereby significantly reducing the size of the sequence optimization problem.

All possible sequences of hydrophilic amino acids (D, E, N, Q, K, R, S, T, A, and H) for the 12 surface positions were screened by our design algorithm. The torsional flexibility of the amino acid side chains was accounted for by considering a discrete set of all allowed conformers of each side chain, called rotamers (Ponder, *et al.*, (1987((supra); Dunbrack, *et al.*, Struc. Biol. Vol. **1**(5):334-340 (1994)). Optimizing the 12 **b**, **c**, and **f** positions each with 10 possible amino acids results in 10^{12} possible sequences which corresponds to ~10^{28} rotamer sequences when using the Dunbrack and Karplus backbone-dependent rotamer library. The immense search problem presented by rotamer sequence optimization is overcome by application of the Dead-End Elimination (DEE) theorem (Desmet, *et al.*, (1992((supra); Desmet, *et al.*, (1994) (supra); Goldstein, (1994) (supra)). Our implementation of the DEE theorem extends its utility to sequence design and rapidly finds the globally optimal sequence in its optimal conformation.

We examined three potential-energy functions for their effectiveness in scoring surface sequences. Each candidate scoring function was used to design the **b**, **c**, and **f** positions of the model coiled coil and the resulting peptide was synthesized and characterized to assess design performance. A hydrogen-bond potential was used to check if predicted hydrogen bonds can contribute to designed protein stability, as expected from studies of hydrogen bonding in proteins and peptides (Stickle, *et al.*, supra; Huyghues-Despointes, *et al.*, supra). Optimizing sequences for hydrogen bonding, however, often buries polar protons that are not involved in hydrogen bonds. This uncompensated

loss of potential hydrogen-bond donors to water prompted examination of a second scoring scheme consisting of a hydrogen-bond potential in conjunction with a penalty for burial of polar protons (Eisenberg, (1986) (supra)). We tested a third scoring scheme which augments the hydrogen bond potential with the empirically derived helix propensities of Baldwin and coworkers (Chakrabartty, *et al.*, supra). Although the physical basis of helix propensities is unclear, they can have a significant effect on protein stability and can potentially be used to improve protein designs (O'Neil & DeGrado, 1990; Zhang, *et al.*, Biochem. 30:2012 (1991); Blaber, *et al.*, Science 260:1637 (1993); O'shea, *et al.*, 1993; Villegas, *et al.*, Folding and Design 1:29 (1996)). A van der Waals potential was used in all cases to account for packing interactions and excluded volume.

Several other sequences for the **b**, **c** and **f** positions were also synthesized and characterized to help discern the relative importance of the hydrogen-bonding and helix-propensity potentials. The sequence designed with the hydrogen-bond potential was randomly scrambled, thereby disrupting the designed interactions but not changing the helix propensity of the sequence. Also, the sequence with the maximum possible helix propensity, all positions set to alanine, was made. Finally, to serve as undesigned controls, the naturally occurring GCN4-p1 sequence and a sequence randomly selected from the hydrophilic amino acid set were synthesized and studied.

Sequence design: Scoring functions and DEE: The protein structure was modeled on the backbone coordinates of GCN4-p1, PDB record 2zta (Bernstein, *et al.*, supra; O'Shea, *et al.*, supra). Atoms of all side chains not optimized were left in their crystallographically determined positions. The program BIOGRAF (Molecular Simulations Incorporated, San Diego, CA) was used to generate explicit hydrogens on the structure which was then conjugate gradient minimized for 50 steps using the DREIDING forcefield (Mayo, *et al.*, 1990, supra). The symmetry of the dimer and lack of interactions of surface residues between the subunits allowed independent design of each subunit. All computations were done using the first monomer to appear in 2zta (chain A). A backbone-dependent rotamer library was used (Dunbrack, *et al.* (1993) (supra)). c_3 angles that were undetermined from the database statistics were assigned the following values: Arg, -60°, 60°, and 180°; Gln, -120°, -60°, 0°, 60°, 120°, and 180°; Glu, 0°, 60°, and 120°; Lys, -60°, 60°, and 180°. c_4 angles that were undetermined from the database statistics were assigned the following values: Arg, -120°, -60°, 60°, 120°, and 180°; Lys, -60°, 60°, and 180°. Rotamers with combinations of c_3 and c_4 that resulted in sequential g^+/g^- or g^-/g^+ angles were eliminated. Uncharged His rotamers were used. A Lennard-Jones 12-6 potential with van der Waals radii scaled by 0.9 (Dahiyat, *et al.*, First fully automatic design of a protein achieved by Caltech scientists, new press release (1997) was used for van der Waals interactions. The hydrogen bond potential consisted of a distance-dependent term and an angle-dependent term, as depicted in Equation 9, above. This hydrogen bond potential is based on the potential used in DREIDING, with more restrictive angle-dependent terms to limit the occurrence of unfavorable hydrogen bond geometries. The angle term varies depending on the hybridization state of the donor and acceptor, as shown in Equations 10 to 13, above.

In Equations 10-13, θ is the donor-hydrogen-acceptor angle, φ is the hydrogen-acceptor-base angle (the base is the atom attached to the acceptor, for example the carbonyl carbon is the base for a carbonyl oxygen acceptor), and φ is the angle between the normals of the planes defined by the six atoms attached to the sp^2 centers (the supplement of φ is used when φ is less than 90°). The hydrogen-bond function is only evaluated when 2.6 Å < *R* < 3.2 Å, φ > 90°, *f* - 109.5° < 90° for the sp^3 donor - sp^3 acceptor case, and, φ > 90° for the sp^3 donor - sp^2 acceptor case; no switching functions were used. Template donors and acceptors that were involved in template-template hydrogen bonds were not included in the donor and acceptor lists. For the purpose of exclusion, a template-template hydrogen bond was considered to exist when 2.5 Å ≤ *R* ≤ 3.3 Å and θ ≤ 135°. A penalty of 2 kcal/mol for polar hydrogen burial, when used, was only applied to buried polar hydrogens not involved in hydrogen bonds, where a hydrogen bond was considered to exist when E_{HB} was less than -2 kcal/mol. This penalty was not applied to template hydrogens. The hydrogen-bond potential was also supplemented with a weak coulombic term that included a distance-dependent dielectric constant of *40R*, where *R* is the interatomic distance. Partial atomic charges were only applied to polar functional groups. A net formal charge of +1 was used for Arg and Lys and a net formal charge of -1 was used for Asp and Glu. Energies associated with α-helical propensities were calculated using equation 14, above. In Equation 14, E_α is the energy of α-helical propensity, $\Delta G°_{aa}$ is the standard free energy of helix propagation of the amino acid, and $\Delta G°_{ala}$ is the standard free energy of helix propagation of alanine used as a standard, and N_{ss} is the propensity scale factor which was set to 3.0. This potential was selected in order to scale the propensity energies to a similar range as the other terms in the scoring function. The DEE optimization followed the methods of our previous work (Dahiyat, *et al.*, (1996) (supra)). Calculations were performed on either a 12 processor, R10000-based Silicon Graphics Power Challenge or a 512 node Intel Delta.

Peptide synthesis and purification and CD analysis was as in Example 1. NMR samples were prepared in 90/10 H_2O/D_2O and 50 mM sodium phosphate buffer at pH 7.0. Spectra were acquired on a Varian Unityplus 600 MHz spectrometer at 25 °C. 32 transients were acquired with 1.5 seconds of solvent presaturation used for water suppression. Samples were ~1 mM. Size exclusion chromatography was performed with a PolyLC hydroxyethyl A column (20 cm x 9 mm) at pH 7.0 in 50 mM phosphate and 150 mM NaCl at 0 °C. GCN4-p1 and p-LI (Harbury, *et al.*, supra) were used as size standards for dimer and tetramer, respectively. 5 µl injections of ~1 mM peptide solution were chromatographed at 0.50 ml/min and monitored at 214 nm. Samples were run in triplicate.

The surface sequences of all of the peptides examined in this study are shown in Table 4.

Table 4. Sequences and properties of the synthesized peptides

Peptide	Design method	Surface Sequence bcf bcf bcf bcf	T_m (°C)	ΔG° (kcal/mol)	N
GCN4-p1	none	KQD EES YHN ARK (SEQ ID NO:31)	57	3.831	2
6A	HB	EKD RER RRE RRE (SEQ ID NO:32)	71	2.193	2
6B	HB + PB	EKQ KER ERE ERQ (SEQ ID NO:33)	72	2.868	2

6C	HB + HP	ARA AAA RRR ARA (SEQ ID NO:34)	69	-2.041	2
6D	scrambled HB	REE RRR EDR KRE (SEQ ID NO:35)	71	2.193	2
6E	random polar	NTR AKS ANH NTQ (SEQ ID NO:36)	15	4.954	2
6F	poly(Ala)	AAA AAA AAA AAA (SEQ ID NO:37)	73	-3.096	4

For clarity only the designed surface residues are shown and they are grouped by position (**b**, **c**, and **f**). The sequence numbers of the designed positions are: 3, 4, 7, 10, 11, 14, 17, 18, 21, 24, 25, and 28. Melting temperatures (T_m's) were determined by circular dichroism and oligomerization states (N) were determined by size exclusion chromatography. $\Delta G°$ is the sum of the standard free energy of helix propagation of the 12 **b**, **c**, and **f** positions (Chakrabartty, *et al.*, 1994). Abbreviations for design methods are: hydrogen bonds (HB), polar hydrogen burial penalty (PB), and helix propensity (HP).

Sequence 6A (SEQ ID NO:32), designed with a hydrogen-bond potential, has a preponderance of Arg and Glu residues that are predicted to form numerous hydrogen bonds to each other. These long chain amino acids are favored because they can extend across turns of the helix to interact with each other and with the backbone. When the optimal geometry of the scrambled 6A (SEQ ID NO:32) sequence, 6D (SEQ ID NO:35), was found with DEE, far fewer hydrogen bonding interactions were present and its score was much worse than 6A's (SEQ ID NO:32). 6B (SEQ ID NO:33), designed with a polar hydrogen burial penalty in addition to a hydrogen-bond potential, is still dominated by long residues such as Lys, Glu and Gln but has fewer Arg. Because Arg has more polar hydrogens than the other amino acids, it more often buries nonhydrogen-bonded protons and therefore is disfavored when using this potential function. 6C (SEQ ID NO:34) was designed with a hydrogen-bond potential and helix propensity in the scoring function and consists entirely of Ala and Arg residues, the amino acids with the highest helix propensities (Chakrabartty, *et al.*, supra). The Arg residues form hydrogen bonds with Glu residues at nearby **e** and **g** positions. The random hydrophilic sequence, 6E (SEQ ID NO:36), possesses no hydrogen bonds and scores very poorly with all of the potential functions used.

The secondary structures and thermal stabilities of the peptides were assessed by circular dichroism (CD) spectroscopy. The CD spectra of the peptides at 1 °C and 40 μM are characteristic of α helices, with minima at 208 and 222 nm, except for the random surface sequence peptide 6E (SEQ ID NO:36). 6E (SEQ ID NO:36) has a spectrum suggestive of a mixture of α helix and random coil with a $[\theta]_{222}$ of -12000 deg cm^2/dmol, while all the other peptides are greater than 90% helical with $[\theta]_{222}$ of less than -30000 deg cm^2/dmol. The melting temperatures (T_m's) of the designed peptides are 12-16 °C higher than the T_m of GCN4-p1 (SEQ ID NO: 31), with the exception of 6E (SEQ ID NO: 36) which has a T_m of 15 °C. CD spectra taken before and after melts were identical indicating reversible thermal denaturation. The redesign of surface positions of this coiled coil produces structures that are much more stable than wildtype GCN4-p1 (SEQ ID NO:31), while a random hydrophilic sequence largely disrupts the peptide's stability.

ize exclusion chromatography (SEC) showed that all the peptides were dimers except for 6F, the all Ala surface sequence, which migrated as a tetramer. These data show that surface redesign did not change the tertiary structure of these peptides, in contrast to some core redesigns (Harbury, *et al.*,

supra). In addition, nuclear magnetic resonance (NMR) spectra of the peptides at ~1 mM showed chemical shift dispersion similar to GCN4-p1 (SEQ ID NO:31) (data not shown).

Peptide 6A (SEQ ID NO:32), designed with a hydrogen-bond potential, melts at 71 °C versus 57 °C for GCN4-p1 (SEQ ID NO: 31), demonstrating that rational design of surface residues can produce structures that are markedly more stable than naturally occurring coiled coils. This gain in stability is probably not due to improved hydrogen bonding since 6D (SEQ ID NO: 35), which has the same surface amino acid composition as 6A (SEQ ID NO: 32) but a scrambled sequence and no predicted hydrogen bonds, also melts at 71 °C. Further, 6B (SEQ ID NO:33) was designed with a different scoring function and has a different sequence and set of predicted hydrogen bonds but a very similar T_m of 72 °C.

An alternative explanation for the increased stability of these sequences relative to GCN4-p1 (SEQ ID NO:31) is their higher helix propensity. The long polar residues selected by the hydrogen bond potential, Lys, Glu, Arg and Gln, are also among the best helix formers (Chakrabartty, *et al.*, supra). Since the effect of helix propensity is not as dependent on sequence position as that of hydrogen bonding, especially far from the helix ends, little effect would be expected from scrambling the sequence of 6A (SEQ ID NO: 32). A rough measure of the helix propensity of the surface sequences, the sum of the standard free energies of helix propagation ($\Sigma\Delta G°$) (Chakrabartty, *et al.*, supra), corresponds to the peptides' thermal stabilities (Table 4). Though $\Sigma\Delta G°$ matches the trend in peptide stability, it is not quantitatively correlated to the increased stability of these coiled coils.

Peptide 6C (SEQ ID NO: 34) was designed with helix propensity as part of the scoring function and it has a $\Sigma\Delta G°$ of -2.041 kcal/mol. Though 6C (SEQ ID NO:34) is more stable than GCN4-p1 (SEQ ID No:31), its T_m of 69 °C is slightly lower than 6A (SEQ ID NO: 32) and 6B (SEQ ID NO:33), in spite of 6C's (SEQ ID NO:34) higher helix propensity. Similarly, 6F has the highest helix propensity possible with an all Ala sequence and a $\Sigma\Delta G°$ of -3.096 kcal/mol, but its T_m of 73 °C is only marginally higher than that of 6A (SEQ ID NO:32) or 6B (SEQ ID NO: 33). 6F also migrates as a tetramer during SEC, not a dimer, likely because its poly(Ala) surface exposes a large hydrophobic patch that could mediate association. Though the results for 6C (SEQ ID NO:34) and 6F (SEQ ID NO:37) support the conclusion that helix propensity is important for surface design, they point out possible limitations in using propensity exclusively. Increasing propensity does not necessarily confer the greatest stability on a structure, perhaps because other factors are being effected unfavorably. Also, as is evident from 6F (SEQ ID NO: 37), changes in the tertiary structure of the protein can occur.

The characterization of these peptides clearly shows that surface residues have a dramatic impact on the stability of α-helical coiled coils. The wide range of stabilities displayed by the different surface designs is notable, with greater than a 50 °C spread between the random hydrophilic sequence (T_m 15 °C) and the designed sequences (T_m 69 - 72 °C). This result is consistent with studies on other proteins that demonstrated the importance of solvent exposed residues (O'Neil & DeGrado, 1990; Zhang, *et al.*, 1991; Minor, *et al.*, (1994) (supra); Smith, *et al.*, Science **270**:980-982 (1995)). Further, these designs have significantly higher T_m's than the wildtype GCN4-p1 sequence, demonstrating that

surface residues can be used to improve stability in protein design (O'shea, *et al.*, supra). Though helix propensity appears to be more important than hydrogen bonding in stabilizing the designed coiled coils, hydrogen bonding could be important in the design and stabilization of other types of secondary structure.

Example 3

Design of a protein containing core, surface and boundary residues using van der Waals, H-bonding, secondary structure and solvation scoring functions

In this example, core, boundary and surface residue work was combined. In selecting a motif to test the integration of our design methodologies, we sought a protein fold that would be small enough to be both computationally and experimentally tractable, yet large enough to form an independently folded structure in the absence of disulfide bonds or metal binding sites. We chose the ββα motif typified by the zinc finger DNA binding module (Pavletich, *et al.* (1991) (supra)). Though it consists of less than 30 residues, this motif contains sheet, helix, and turn structures. Further, recent work by Imperiali and coworkers who designed a 23 residue peptide, containing an unusual amino acid (D-proline) and a non-natural amino acid (3-(1,10-phenanthrol-2-yl)-L-alanine), that takes this structure has demonstrated the ability of this fold to form in the absence of metal ions (Struthers, *et al.*, 1996a). The Brookhaven Protein Data Bank (PDB) (Bernstein, *et al.*, 1977) was examined for high resolution structures of the ββα motif, and the second zinc finger module of the DNA binding protein Zif268 (PDB code 1zaa) was selected as our design template (Pavletich, *et al.* (1991) (supra)). The backbone of the second module aligns very closely with the other two zinc fingers in Zif268 and with zinc fingers in other proteins and is therefore representative of this fold class. 28 residues were taken from the crystal structure starting at lysine 33 in the numbering of PDB entry 1zaa which corresponds to our position 1. The first 12 residues comprise the β sheet with a tight turn at the 6^{th} and 7^{th} positions. Two residues connect the sheet to the helix, which extends through position 26 and is capped by the last two residues.

In order to assign the residue positions in the template structure into core, surface or boundary classes, the extent of side-chain burial in Zif268 and the direction of the Cα-Cβ vectors were examined. The small size of this motif limits to one (position 5) the number of residues that can be assigned unambiguously to the core while six residues (positions 3, 12, 18, 21, 22, and 25) were classified as boundary. Three of these residues are from the sheet (positions 3, 5, and 12) and four are from the helix (positions 18, 21, 22, and 25). One of the zinc binding residues of Zif268 is in the core and two are in the boundary, but the fourth, position 8, has a Cα-Cβ vector directed away from the protein's geometric center and is therefore classified as a surface position. The other surface positions considered by the design algorithm are 4, 9, and 11 from the sheet, 15, 16, 17, 19, 20, and 23 from the helix and 14, 27, and 28 which cap the helix ends. The remaining exposed positions, which either were in turns, had irregular backbone dihedrals or were partially buried, were not included in the sequence selection for this initial study. As in our previous studies, the amino acids considered at the core positions during sequence selection were A, V, L, I, F, Y, and W; the amino

acids considered at the surface positions were A, S, T, H, D, N, E, Q, K, and R; and the combined core and surface amino acid sets (16 amino acids) were considered at the boundary positions.

In total, 20 out of 28 positions of the template were optimized during sequence selection. The algorithm first selects Gly for all positions with φ angles greater than 0° in order to minimize backbone strain (residues 9 and 27). The 18 remaining residues were split into two sets and optimized separately to speed the calculation. One set contained the 1 core, the 6 boundary positions and position 8 which resulted in 1.2×10^9 possible amino acid sequences corresponding to 4.3×10^{19} rotamer sequences. The other set contained the remaining 10 surface residues which had 10^{10} possible amino acid sequences and 4.1×10^{23} rotamer sequences. The two groups do not interact strongly with each other making their sequence optimizations mutually independent, though there are strong interactions within each group. Each optimization was carried out with the non-optimized positions in the template set to the crystallographic coordinates.

The optimal sequences found from the two calculations were combined and are shown in Figure 8 (SEQ ID NOS:1 and 2) aligned with the sequence from the second zinc finger of Zif268 (SEQ ID NO:1). Even though all of the hydrophilic amino acids were considered at each of the boundary positions, only nonpolar amino acids were selected. The calculated seven core and boundary positions form a well-packed buried cluster. The Phe side chains selected by the algorithm at the zinc binding His positions, 21 and 25, are 80% buried and the Ala at 5 is 100% buried while the Lys at 8 is greater than 60% exposed to solvent. The other boundary positions demonstrate the strong steric constraints on buried residues by packing similar side chains in an arrangement similar to Zif268. The calculated optimal configuration buried ~830 $Å^2$ of nonpolar surface area, with Phe 12 (96% buried) and Leu 18 (88% buried) anchoring the cluster. On the helix surface, the algorithm positions Asn 14 as a helix N-cap with a hydrogen bond between its side-chain carbonyl oxygen and the backbone amide proton of residue 16. The six charged residues on the helix form three pairs of hydrogen bonds, though in our coiled coil designs helical surface hydrogen bonds appeared to be less important than the overall helix propensity of the sequence. Positions 4 and 11 on the exposed sheet surface were selected to be Thr, one of the best β-sheet forming residues (Kim & Berg, 1993; Minor, *et al.*, (1994) (supra); Smith, *et al.*, (1995) (supra)).

Combining the 20 designed positions with the Zif268 (SEQ ID NO:1) amino acids at the remaining 8 sites results in a peptide with overall 39% (11/28) homology to Zif268, which reduces to 15% (3/20) homology when only the designed positions are considered. A BLAST (Altschul, *et al.*, 1990) search of the non-redundant protein sequence database of the National Center for Biotechnology Information finds weak homology, less than 40%, to several zinc finger proteins and fragments of other unrelated proteins. None of the alignments had significance values less than 0.26. By objectively selecting 20 out of 28 residues on the Zif268 (SEQ ID NO:1) template, a peptide with little homology to known proteins and no zinc binding site was designed.

Experim ntal characterization: The far UV circular dichroism (CD) spectrum of the designed molecule, pda8d, shows a maximum at 195 nm and minima at 218 nm and 208 nm, which is

indicative of a folded structure. The thermal melt is weakly cooperative, with an inflection point at 39 °C, and is completely reversible. The broad melt is consistent with a low enthalpy of folding which is expected for a motif with a small hydrophobic core. This behavior contrasts the uncooperative transitions observed for other short peptides (Weiss & Keutmann, 1990; Scholtz, *et al.*, PNAS USA 88:2854 (1991); Struthers, *et al.*, J. Am. Chem. Soc. 118:3073 (1996b)).

Sedimentation equilibrium studies at 100 μM and both 7 °C and 25 °C give a molecular mass of 3490, in good agreement with the calculated mass of 3362, indicating the peptide is monomeric. At concentrations greater than 500 μM, however, the data do not fit well to an ideal single species model. When the data were fit to a monomer-dimer- tetramer model, dissociation constants of 0.5 - 1.5 mM for monomer-to-dimer and greater than 4 mM for dimer-to-tetramer were found, though the interaction was too weak to accurately measure these values. Diffusion coefficient measurements using the water-sLED pulse sequence (Altieri, *et al.*, 1995) agreed with the sedimentation results: at 100 μM pda8d has a diffusion coefficient close to that of a monomeric zinc finger control, while at 1.5 mM the diffusion coefficient is similar to that of protein G β1, a 56 residue protein. The CD spectrum of pda8d is concentration independent from 10 μM to 2.6 mM. NMR COSY spectra taken at 2.1 mM and 100 μM were almost identical with 5 of the Hα-HN crosspeaks shifted no more than 0.1 ppm and the rest of the crosspeaks remaining unchanged. These data indicate that pda8d undergoes a weak association at high concentration, but this association has essentially no effect on the peptide's structure.

The NMR chemical shifts of pda8d are well dispersed, suggesting that the protein is folded and well-ordered. The Hα-HN fingerprint region of the TOCSY spectrum is well-resolved with no overlapping resonances (Figure (9A) and all of the Hα and HN resonances have been assigned. NMR data were collected on a Varian Unityplus 600 MHz spectrometer equipped with a Nalorac inverse probe with a self-shielded z-gradient. NMR samples were prepared in 90/10 H_2O/D_2O or 99.9% D_2O with 50 mM sodium phosphate at pH 5.0. Sample pH was adjusted using a glass electrode with no correction for the effect of D_2O on measured pH. All spectra for assignments were collected at 7 □C. Sample concentration was approximately 2 mM. NMR assignments were based on standard homonuclear methods using DQF-COSY, NOESY and TOCSY spectra (Wuthrich, NMR of Proteins and Nucleic Acids (John Wiley & Sons, New York, 1986). NOESY and TOCSY spectra were acquired with 2K points in F2 and 512 increments in F1 and DQF-COSY spectra were acquired with 4K points in F2 and 1024 increments in F1. All spectra were acquired with a spectral width of 7500 Hz and 32 transients. NOESY spectra were recorded with mixing times of 100 and 200 ms and TOCSY spectra were recorded with an isotropic mixing time of 80 ms. In TOCSY and DQF-COSY spectra water suppression was achieved by presaturation during a relaxation delay of 1.5 and 2.0 s, respectively. Water suppression in the NOESY spectra was accomplished with the WATERGATE pulse sequence (Piotto, *et al.*, 1992). Chemical shifts were referenced to the HOD resonance. Spectra were zero-filled in both F2 and F1 and apodized with a shifted gaussian in F2 and a cosine bell in F1 (NOESY and TOCSY) or a 30° shifted sine bell in F2 and a shifted gaussian in F1 (DQF-COSY).

Water-sLED experiments (Altieri, *et al.*, 1995) were run at 25 °C at 1.5 mM, 400 μM and 100 μM in 99.9% D_2O with 50 mM sodium phosphate at pH 5.0. Axial gradient field strength was varied from 3.26 to 53.1 G/cm and a diffusion time of 50 ms was used. Spectra were processed with 2 Hz line broadening and integrals of the aromatic and high field aliphatic protons were calculated and fit to an equation relating resonance amplitude to gradient strength in order to extract diffusion coefficients (Altieri, *et al.*, 1995). Diffusion coefficients were 1.48×10^{-7}, 1.62×10^{-7} and 1.73×10^{-7} cm^2/s at 1.5 mM, 400 μM and 100 μM, respectively. The diffusion coefficient for the zinc finger monomer control was 1.72×10^{-7} cm^2/s and for protein G b1 was 1.49×10^{-7} cm^2/s.

All unambiguous sequential and medium-range NOEs are shown in Figure 9A. Hα-HN and/or HN-HN NOEs were found for all pairs of residues except R6-I7 and K16-E17, both of which have degenerate HN chemical shifts, and P2-Y3 which have degenerate Hα chemical shifts. An NOE is present, however, from a P2 Hδ to the Y3 HN analogous to sequential HN-HN connections. Also, strong K1 Hα to P2 Hδ NOEs are present and allowed completion of the resonance assignments.

The structure of pda8d was determined using 354 NOE restraints (12.6 restraints per residue) that were non-redundant with covalent structure. An ensemble of 32 structures (data not shown) was obtained using X-PLOR (Brunger, 1992) with standard protocols for hybrid distance geometry-simulated annealing. The structures in the ensemble had good covalent geometry and no NOE restraint violations greater than 0.3 Å. As shown in Table 5, the backbone was well defined with a root-mean-square (rms) deviation from the mean of 0.55 Å when the disordered termini (residues 1, 2, 27, and 28) were excluded. The rms deviation for the backbone (3-26) plus the buried side chains (residues 3, 5, 7, 12, 18, 21, 22, and 25) was 1.05 Å.

Table 5. NMR structure determination of pda8d: distance restraints, structural statistics, atomic root-mean-square (rms) deviations, and comparison to the design target. <SA> are the 32 simulated annealing structures, SA is the average structure and SD is the standard deviation. The design target is the backbone of Zif268.

Distance restraints

Intraresidue	148
Sequential	94
Short range (|i-j| = 2-5 residues)	78
Long range (|i-j| > 5 residues)	34
Total	354

Structural statistics

	<SA> ± SD
Rms deviation from distance restraints (Å)	0.049 ±.004
Rms deviation from idealized geometry (Å)	
Bonds (Å)	0.0051 ± 0.0004
Angles (degrees)	0.76 ± 0.04
Impropers (degrees)	0.56 ± 0.04

Atomic rms deviations (Å)*

	<SA> vs. SA ± SD
Backbone	0.55 ± 0.03
Backbone + nonpolar side chains	1.05 ± 0.06
Heavy atoms	1.25 ± 0.04

Atomic rms deviations between pda8d and the design target (Å)*

	SA vs. target
Backbone	1.04
Heavy atoms	2.15

*Atomic rms deviations are for residues 3 to 26, inclusive. The termini, residues 1, 2, 27, and 28, were highly disordered and had very few non-sequential or non-intraresidue contacts.

The NMR solution structure of pda8d shows that it folds into a bba motif with well-defined secondary structure elements and tertiary organization which match the design target. A direct comparison of the design template, the backbone of the second zinc finger of Zif268, to the pda8d solution structure highlights their similarity (data not shown). Alignment of the pda8d backbone to the design target is excellent, with an atomic rms deviation of 1.04 Å (Table 5). Pda8d and the design target correspond throughout their entire structures, including the turns connecting the secondary structure elements.

In conclusion, the experimental characterization of pda8d shows that it is folded and well-ordered with a weakly cooperative thermal transition, and that its structure is an excellent match to the design target. To our knowledge, pda8d is the shortest sequence of naturally occurring amino acids that folds to a unique structure without metal binding, oligomerization or disulfide bond formation (McKnight, *et al.*, Nature Struc. Biol. 4:180 (1996)). The successful design of pda8d supports the use of objective, quantitative sequence selection algorithms for protein design. This robustness suggests that the program can be used to design sequences for de novo backbones.

Example 4

Protein design using a scaled van der Waals scoring function in the core region

An ideal model system to study core packing is the β1 immunoglobulin-binding domain of streptococcal protein G (Gβ1) (Gronenborn, *et al.*, Science 253:657 (1991); Alexander, *et al.*, Biochem. 31: 3597 (1992); Barchi, *et al.*, Protein Sci. 3:15 (1994); Gallagher, *et al.*, 1994; Kuszewski, *et al.*, 1994; Orban, *et al.*, 1995). Its small size, 56 residues, renders computations and experiments tractable. Perhaps most critical for a core packing study, Gβ1 contains no disulfide bonds and does not require a cofactor or metal ion to fold. Further, Gβ1 contains sheet, helix and turn structures and is without the repetitive side-chain packing patterns found in coiled coils or some helical bundles. This lack of periodicity reduces the bias from a particular secondary or tertiary structure and necessitates the use of an objective side-chain selection program to examine packing effects.

Sequence positions that constitute the core were chosen by examining the side-chain solvent accessible surface area of Gβ1. Any side chain exposing less than 10% of its surface was considered buried. Eleven residues meet this criteria, with seven from the β sheet (positions 3, 5, 7, 20, 43, 52 and 54), three from the helix (positions 26, 30, and 34) and one in an irregular secondary structure (position 39). These positions form a contiguous core. The remainder of the protein structure, including all other side chains and the backbone, was used as the template for sequence selection calculations at the eleven core positions.

All possible core sequences consisting of alanine, valine, leucine, isoleucine, phenylalanine, tyrosine or tryptophan (A, V, L, I, F, Y or W) were considered. Our rotamer library was similar to that used by Desmet and coworkers (Desmet, *et al.*, (1992) (supra)). Optimizing the sequence of the core of Gβ1 (SEQ ID NO:38) with 217 possible hydrophobic rotamers at all 11 positions results in 217^{11}, or $5x10^{25}$, rotamer sequences. Our scoring function consisted of two components: a van der Waals energy term and an atomic solvation term favoring burial of hydrophobic surface area. The van der Waals radii of all atoms in the simulation were scaled by a factor α (Eqn. 3) to change the importance of packing effects. Radii were not scaled for the buried surface area calculations. By predicting core sequences with various radii scalings and then experimentally characterizing the resulting proteins, a rigorous study of the importance of packing effects on protein design is possible.

The protein structure was modeled on the backbone coordinates of Gβ1, PDB record 1pga (Bernstein, *et al.*, supra; Gallagher, *et al.*, 1994). Atoms of all side chains not optimized were left in

their crystallographically determined positions. The program BIOGRAF (Molecular Simulations Incorporated, San Diego, CA) was used to generate explicit hydrogens on the structure which was then conjugate gradient minimized for 50 steps using the Dreiding forcefield (Mayo, *et al.*, 1990, supra). The rotamer library, DEE optimization and Monte Carlo search was as outlined above. A Lennard-Jones 12-6 potential was used for van der Waals interactions, with atomic radii scaled for the various cases as discussed herein. The Richards definition of solvent-accessible surface area (Lee & Richards, supra) was used and areas were calculated with the Connolly algorithm (Connolly, (1983) (supra)). An atomic solvation parameter, derived from our previous work, of 23 cal/mol/Å^2 was used to favor hydrophobic burial and to penalize solvent exposure. To calculate side-chain nonpolar exposure in our optimization framework, we first consider the total hydrophobic area exposed by a rotamer in isolation. This exposure is decreased by the area buried in rotamer/template contacts, and the sum of the areas buried in pairwise rotamer/rotamer contacts.

Global optimum sequences for various values of the radius scaling factor α were found using the Dead-End Elimination theorem (Table 6) (SEQ ID NOS:38 – 49). Optimal sequences, and their corresponding proteins, are named by the radius scale factor used in their design. For example, the sequence designed with a radius scale factor of α = 0.90 is called α90 (SEQ ID NO:43).

Table 6.

Gβ1 sequence

α	vol	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	TRP	PHE	VAL	(SEQ ID NO:38)
		3	5	7	20	26	30	34	39	43	52	54	
0.70	1.28	TRP	TYR	ILE	ILE	PHE	TRP	LEU	ILE	PHE	LEU	ILE	(SEQ ID NO:39)
0.75	1.23	PHE	ILE	PHE	ILE	VAL	TRP	VAL	LEU			ILE	(SEQ ID NO:40)
0.80	1.13	PHE		ILE				ILE	ILE		TRP	ILE	(SEQ ID NO:41)
0.85	1.15	PHE		ILE				LEU	ILE		TRP	PHE	(SEQ ID NO:42)
0.90	1.01	PHE		ILE					ILE				(SEQ ID NO:43)
0.95	1.01	PHE		ILE					ILE				(SEQ ID NO:44)
1.0	0.99	PHE		VAL					ILE				(SEQ ID NO:45)
1.05	0.93	PHE		ALA									(SEQ ID NO:46)
1.075	0.83	ALA	ALA	ILE			ILE				ILE	ILE	(SEQ ID NO:47)
1.10	0.77	ALA		ALA			ALA				ILE	ILE	(SEQ ID NO:48)
1.15	0.68	ALA	ALA	ALA			ALA				LEU		(SEQ ID NO:49)

In Table 6, the Gβ1 sequence (SEQ ID NO:38) and position numbers are shown at the top. vol is the fracton of core side-chain volume relative to the Gβ1 sequence (SEQ ID NO:38). A vertical bar indicates identity with the Gβ1 sequence (SEQ ID NO:38).

α100 was designed with α = 1.0 and hence serves as a baseline for full incorporation of steric effects. The α100 sequence (SEQ ID NO:45) is very similar to the core sequence of (Gβ1 (SEQ ID NO:38) (Table 6) even though no information about the naturally occurring sequence was used in the side-chain selection algorithm. Variation of α from 0.90 to 1.05 caused little change in the optimal sequence, demonstrating the algorithm's robustness to minor parameter perturbations. Further, the packing arrangements predicted with α = 0.90 - 1.05 closely match Gβ1 with average χ angle differences of only 4° from the crystal structure. The high identity and conformational similarity to Gβ1 imply that, when packing constraints are used, backbone conformation strongly determines a single family of well packed core designs. Nevertheless, the constraints on core packing were being modulated by α as demonstrated by Monte Carlo searches for other low energy sequences. Several alternate sequences and packing arrangements are in the twenty best sequences found by the Monte Carlo procedure when α = 0.90. These alternate sequences score much worse when α = 0.95, and when α = 1.0 or 1.05 only strictly conservative packing geometries have low energies. Therefore, α = 1.05 and α = 0.90 define the high and low ends, respectively, of a range where packing specificity dominates sequence design.

For α < 0.90, the role of packing is reduced enough to let the hydrophobic surface potential begin to dominate, thereby increasing the size of the residues selected for the core (Table 6). A significant change in the optimal sequence appears between α = 0.90 and 0.85 with both α85 and α80 containing three additional mutations relative to α90. Also, α85 and α80 have a 15% increase in total side-chain volume relative to Gβ1. As α drops below 0.80 an additional 10% increase in side-chain volume and numerous mutations occur, showing that packing constraints have been overwhelmed by the drive to bury nonpolar surface. Though the jumps in volume and shifts in packing arrangement appear to occur suddenly for the optimal sequences, examination of the suboptimal low energy sequences by Monte Carlo sampling demonstrates that the changes are not abrupt. For example, the α85 optimal sequence is the 11th best sequence when α = 0.90, and similarly, the α90 optimal sequence is the 9th best sequence when α = 0.85.

For α > 1.05 atomic van der Waals repulsions are so severe that most amino acids cannot find any allowed packing arrangements, resulting in the selection of alanine for many positions. This stringency is likely an artifact of the large atomic radii and does not reflect increased packing specificity accurately. Rather, α = 1.05 is the upper limit for the usable range of van der Waals scales within our modeling framework.

Experimental characterization of core designs. Variation of the van der Waals scale factor a results in four regimes of packing specificity: regime 1 where 0.9 ≤ α ≤ 1.05 and packing constraints dominate the sequence selection; regime 2 where 0.8 ≤ α < 0.9 and the hydrophobic solvation potential begins to compete with packing forces; regime 3 where α < 0.8 and hydrophobic solvation dominates the design; and, regime 4 where α > 1.05 and van der Waals repulsions appear to be too severe to allow meaningful sequence selection. Sequences that are optimal designs were selected from each of the regimes for synthesis and characterization. They are α 90 from regime 1, α 85 from

regime 2, α 70 from regime 3 and α 107 from regime 4. For each of these sequences, the calculated amino acid identities of the eleven core positions are shown in Table 6; the remainder of the protein sequence matches Gβ1. The goal was to study the relation between the degree of packing specificity used in the core design and the extent of native-like character in the resulting proteins.

Peptide synthesis and purification. With the exception of the eleven core positions designed by the sequence selection algorithm, the sequences synthesized match Protein Data Bank entry 1pga. Peptides were synthesized using standard Fmoc chemistry, and were purified by reverse-phase HPLC. Matrix assisted laser desorption mass spectrometry found all molecular weights to be within one unit of the expected masses.

CD and fluorescence spectroscopy and size exclusion chromatography. The solution conditions for all experiments were 50 mM sodium phosphate buffer at pH 5.5 and 25 °C unless noted. Circular dichroism spectra were acquired on an Aviv 62DS spectrometer equipped with a thermoelectric unit. Peptide concentration was approximately 20 μM. Thermal melts were monitored at 218 nm using 2° increments with an equilibration time of 120 s. T_m's were defined as the maxima of the derivative of the melting curve. Reversibility for each of the proteins was confirmed by comparing room temperature CD spectra from before and after heating. Guanidinium chloride denaturation measurements followed published methods (Pace, Methods. Enzymol. 131:266 (1986)). Protein concentrations were determined by UV spectrophotometry. Fluorescence experiments were performed on a Hitachi F-4500 in a 1 cm pathlength cell. Both peptide and ANS concentrations were 50 μM. The excitation wavelength was 370 nm and emission was monitored from 400 to 600 nm. Size exclusion chromatography was performed with a PolyLC hydroxyethyl A column at pH 5.5 in 50 mM sodium phosphate at 0 °C. Ribonuclease A, carbonic anhydrase and Gβ1 were used as molecular weight standards. Peptide concentrations during the separation were ~15 μM as estimated from peak heights monitored at 275 nm.

Nuclear magnetic resonance spectroscopy. Samples were prepared in 90/10 H_2O/D_2O and 50 mM sodium phosphate buffer at pH 5.5. Spectra were acquired on a Varian Unityplus 600 MHz spectrometer at 25 °C. Samples were approximately 1 mM, except for α70 which had limited solubility (100 μM). For hydrogen exchange studies, an NMR sample was prepared, the pH was adjusted to 5.5 and a spectrum was acquired to serve as an unexchanged reference. This sample was lyophilized, reconstituted in D_2O and repetitive acquisition of spectra was begun immediately at a rate of 75 s per spectrum. Data acquisition continued for ~20 hours, then the sample was heated to 99 °C for three minutes to fully exchange all protons. After cooling to 25 °C, a final spectrum was acquired to serve as the fully exchanged reference. The areas of all exchangeable amide peaks were normalized by a set of non-exchanging aliphatic peaks. pH values, uncorrected for isotope effects, were measured for all the samples after data acquisition and the time axis was normalized to correct for minor differences in pH (Rohl, *et al.*, Biochem. 31:1263 (1992)).

α 90 and α 85 have ellipticities and spectra very similar to Gβ1 (not shown), suggesting that their secondary structure content is comparable to that of Gb1 (Figure 10). Conversely, α 70 has much

weaker ellipticity and a perturbed spectrum, implying a loss of secondary structure relative to Gβ1. α 107 has a spectrum characteristic of a random coil. Thermal melts monitored by CD are shown in Figure 10B. α85 and α 90 both have cooperative transitions with melting temperatures (T_m's) of 83 °C and 92 °C, respectively. α 107 shows no thermal transition, behavior expected from a fully unfolded polypeptide, and α 70 has a broad, shallow transition, centered at ~40 °C, characteristic of partially folded structures. Relative to Gβ1, which has a T_m of 87 °C (Alexander, *et al.*, supra), α 85 is slightly less thermostable and α 90 is more stable. Chemical denaturation measurements of the free energy of unfolding (ΔG_u) at 25 °C match the trend in T_m's.

α 90 has a larger ΔG_u than that reported for Gβ1 (Alexander, *et al.*, supra) while α 85 is slightly less stable. It was not possible to measure ΔG_u for α 70 or α107 because they lack discernible transitions.

The extent of chemical shift dispersion in the proton NMR spectrum of each protein was assessed to gauge each protein's degree of native-like character (data not shown). α 90 possesses a highly dispersed spectrum, the hallmark of a well-ordered native protein. α 85 has diminished chemical shift dispersion and peaks that are somewhat broadened relative to α 90, suggesting a moderately mobile structure that nevertheless maintains a distinct fold. α 70's NMR spectrum has almost no dispersion. The broad peaks are indicative of a collapsed but disordered and fluctuating structure. α 107 has a spectrum with sharp lines and no dispersion, which is indicative of an unfolded protein.

Amide hydrogen exchange kinetics are consistent with the conclusions reached from examination of the proton NMR spectra. Measuring the average number of unexchanged amide protons as a function of time for each of the designed proteins results as follows (data not shown): α 90 protects ~13 protons for over 20 hours of exchange at pH 5.5 and 25 °C. The α 90 exchange curve is indistinguishable from Gβ1's (not shown). α 85 also maintains a well-protected set of amide protons, a distinctive feature of ordered native-like proteins. The number of protected protons, however, is only about half that of α 90. The difference is likely due to higher flexibility in some parts of the α 85 structure. In contrast, α 70 and α 107 were fully exchanged within the three minute dead time of the experiment, indicating highly dynamic structures.

Near UV CD spectra and the extent of 8-anilino-1-naphthalene sulfonic acid (ANS) binding were used to assess the structural ordering of the proteins. The near UV CD spectra of α85 and α90 have strong peaks as expected for proteins with aromatic residues fixed in a unique tertiary structure while α70 and α107 have featureless spectra indicative of proteins with mobile aromatic residues, such as non-native collapsed states or unfolded proteins. α70 also binds ANS well, as indicated by a three-fold intensity increase and blue shift of the ANS emission spectrum. This strong binding suggests that α70 possesses a loosely packed or partially exposured cluster of hydrophobic residues accessible to ANS. ANS binds α85 weakly, with only a 25% increase in emission intensity, similar to the association seen for some native proteins (Semisotnov, *et al.*, Biopolymers 31:119 (1991)). α90 and α107 cause no change in ANS fluorescence. All of the proteins migrated as monomers during size exclusion chromatography.

In summary, α 90 is a well-packed native-like protein by all criteria, and it is more stable than the naturally occurring Gβ1 sequence, possibly because of increased hydrophobic surface burial. α 85 is also a stable, ordered protein, albeit with greater motional flexibility than α90, as evidenced by its NMR spectrum and hydrogen exchange behavior. α70 has all the features of a disordered collapsed globule: a non-cooperative thermal transition, no NMR spectral dispersion or amide proton protection, reduced secondary structure content and strong ANS binding. α107 is a completely unfolded chain, likely due to its lack of large hydrophobic residues to hold the core together. The clear trend is a loss of protein ordering as α decreases below 0.90.

The different packing regimes for protein design can be evaluated in light of the experimental data. In regime 1, with $0.9 \leq \alpha \leq 1.05$, the design is dominated by packing specificity resulting in well-ordered proteins. In regime 2, with $0.8 \leq \alpha < 0.9$, packing forces are weakened enough to let the hydrophobic force drive larger residues into the core which produces a stable well-packed protein with somewhat increased structural motion. In regime 3, $\alpha < 0.8$, packing forces are reduced to such an extent that the hydrophobic force dominates, resulting in a fluctuating, partially folded structure with no stable core packing. In regime 4, $\alpha > 1.05$, the steric forces used to implement packing specificity are scaled too high to allow reasonable sequence selection and hence produce an unfolded protein. These results indicate that effective protein design requires a consideration of packing effects. Within the context of a protein design algorithm, we have quantitatively defined the range of packing forces necessary for successful designs. Also, we have demonstrated that reduced specificity can be used to design protein cores with alternative packings.

To take advantage of the benefits of reduced packing constraints, protein cores should be designed with the smallest α that still results in structurally ordered proteins. The optimal protein sequence from regime 2, α85, is stable and well packed, suggesting $0.8 \leq \alpha < 0.9$ as a good range. NMR spectra and hydrogen exchange kinetics, however, clearly show that α85 is not as structurally ordered as α90. The packing arrangements predicted by our program for W43 in α85 and α90 present a possible explanation. For α90, W43 is predicted to pack in the core with the same conformation as in the crystal structure of Gβ1. In α85, the larger side chains at positions 34 and 54, leucine and phenylalanine respectively, compared to alanine and valine in α90, force W43 to expose 91 $Å^2$ of nonpolar surface compared to 19 $Å^2$ in α90. The hydrophobic driving force this exposure represents seems likely to stabilize alternate conformations that bury W43 and thereby could contribute to α85's conformational flexibility (Dill, 1985; Onuchic, *et al.*, 1996). In contrast to the other core positions, a residue at position 43 can be mostly exposed or mostly buried depending on its side-chain conformation. We designate positions with this characteristic as boundary positions, which pose a difficult problem for protein design because of their potential to either strongly interact with the protein's core or with solvent.

A scoring function that penalizes the exposure of hydrophobic surface area might assist in the design of boundary residues. Dill and coworkers used an exposure penalty to improve protein designs in a theoretical study (Sun, *et al.*, Protein Eng. **8**(12)1205-1213 (1995)).

A nonpolar exposure penalty would favor packing arrangements that either bury large side chains in the core or replace the exposed amino acid with a smaller or more polar one. We implemented a side-chain nonpolar exposure penalty in our optimization framework and used a penalizing solvation parameter with the same magnitude as the hydrophobic burial parameter.

The results of adding a hydrophobic surface exposure penalty to our scoring function are shown in Table 7.

Table 7.

α=0.85

#	A_{np}	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	TRP	PHE	VAL	(SEQ ID NO:38)
		3	5	7	20	26	30	34	39	43	52	54	
1	109	PHE	□	ILE	□	□	□	LEU	ILE	□	TRP	PHE	(SEQ ID NO: 50)
2	109	□	□	ILE	□	□	□	LEU	ILE	□	TRP	PHE	(SEQ ID NO:51)
3	104	PHE	□	ILE	□	□	□	LEU	ILE	□	□	PHE	(SEQ ID NO:52)
4	104	□	□	ILE	□	□	□	LEU	ILE	□	□	PHE	(SEQ ID NO:53)
5	108	PHE	□	ILE	□	□	□	LEU	□	□	TRP	PHE	(SEQ ID NO:54)
6	62	PHE	□	ILE	□	□	□	LEU	ILE	VAL	TRP	PHE	(SEQ ID NO:55)
7	103	PHE	□	ILE	□	□	□	LEU	ILE	□	TYR	PHE	(SEQ ID NO:56)
8	109	PHE	□	VAL	□	□	□	LEU	ILE	□	TRP	PHE	(SEQ ID NO:57)
9	30	PHE	□	ILE	□	□	□	□	ILE	□	□	□	(SEQ ID NO:58)
10	38	PHE	□	ILE	□	□	□	□	ILE	□	TRP	□	(SEQ ID NO:59)
11	108	□	□	ILE	□	□	□	LEU	□	□	TRP	PHE	(SEQ ID NO:60)
12	62	□	□	ILE	□	□	□	LEU	ILE	VAL	TRP	PHE	(SEQ ID NO:61)
13	109	PHE	□	ILE	□	□	TYR	LEU	ILE	□	TRP	PHE	(SEQ ID NO:62)
14	103	□	□	ILE	□	□	□	LEU	ILE	□	TYR	PHE	(SEQ ID NO:63)
15	109	□	□	VAL	□	□	□	LEU	ILE	□	TRP	PHE	(SEQ ID NO:64)

Table 7 depicts the 15 best sequences (SEQ ID NOS:50 –64) for the core positions of Gβ1 (SEQ ID NO:38) using α = 0.85 without an exposure penalty. A_{np} is the exposed nonpolar surface area in $Å^2$.

When α = 0.85 the nonpolar exposure penalty dramatically alters the ordering of low energy sequences. The α85 sequence, the former ground state, drops to 7th and the rest of the 15 best sequences expose far less hydrophobic area because they bury W43 in a conformation similar to α90 (model not shown). The exceptions are the 8th and 14th sequences (SEQ ID NOS: 57 and 63,

respectively), which reduce the size of the exposed boundary residue by replacing W43 with an isoleucine, and the 13th best sequence which replaces W43 with a valine. The new ground state sequence is very similar to α90, with a single valine to isoleucine mutation, and should share α90's stability and structural order. In contrast, when α = 0.90, the optimal sequence does not change and the next 14 best sequences, found by Monte Carlo sampling, change very little. This minor effect is not surprising, since steric forces still dominate for α = 0.90 and most of these sequences expose very little surface area. Burying W43 restricts sequence selection in the core somewhat, but the reduced packing forces for α = 0.85 still produce more sequence variety than α = 0.90. The exposure penalty complements the use of reduced packing specificity by limiting the gross overpacking and solvent exposure that occurs when the core's boundary is disrupted. Adding this constraint should allow lower packing forces to be used in protein design, resulting in a broader range of high-scoring sequences and reduced bias from fixed backbone and discrete rotamers.

To examine the effect of substituting a smaller residue at a boundary position, we synthesized and characterized the 13th best sequence of the α = 0.85 optimization with exposure penalty (Table 8).

Table 8.

α=0.85 exposure p nalty

#	A_{np}	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	TRP	PHE	VAL	(SEQ ID NO:38)
		3	5	7	20	26	30	34	39	43	52	54	
1	30	PHE		ILE					ILE			ILE	(SEQ ID NO:65)
2	29	PHE		ILE				ILE	ILE				(SEQ ID NO:66)
3	29	PHE	ILE	PHE					ILE				(SEQ ID NO:67)
4	30			ILE					ILE			ILE	(SEQ ID NO:68)
5	29			ILE				ILE	ILE				(SEQ ID NO:69)
6	29		ILE	PHE					ILE				(SEQ ID NO:70)
7	109	PHE		ILE				LEU	ILE		TRP	PHE	(SEQ ID NO:71)
8	52	PHE		ILE				LEU	ILE	ILE		PHE	(SEQ ID NO:72)
9	29			ILE					ILE				(SEQ ID NO:73)
10	29	PHE		ILE					ILE				(SEQ ID NO:74)
11	109			ILE				LEU	ILE		TRP	PHE	(SEQ ID NO:75)
12	38	PHE		ILE					ILE		TRP	ILE	(SEQ ID NO:76)
13	62	PHE		ILE				LEU	ILE	VAL	TRP	PHE	(SEQ ID NO:77)
14	52			ILE				LEU	ILE	ILE		PHE	(SEQ ID NO:78)
15	30	PHE		ILE					ILE		TYR	ILE	(SEQ ID NO:79)

Table 8 depicts the 15 best sequences (SEQ ID NOS: 65 – 79) of the core positions of Gβ1 (SEQ ID NO:38) using α = 0.85 with an exposure penalty. A_{np} is the exposed nonpolar surface area in $Å^2$.

This sequence, α85W43V, replaces W43 with a valine but is otherwise identical to α85. Though the 8th and 14th sequences (SEQ ID NOS:72 and 78, respectively) also have a smaller side chain at position 43, additional changes in their sequences relative to α85 would complicate interpretation of the effect of the boundary position change. Also, α85W43V has a significantly different packing arrangement compared to Gβ1, with 7 out of 11 positions altered, but only an 8% increase in side-chain volume. Hence, α85W43V is a test of the tolerance of this fold to a different, but nearly volume conserving, core. The far UV CD spectrum of α85W43V is very similar to that of Gβ1 with an ellipticity at 218 nm of -14000 deg cm^2/dmol. While the secondary structure content of α85W43V is native-like, its T_m is 65 °C, nearly 20 °C lower than α85. In contrast to α85W43V's decreased stability, its NMR spectrum has greater chemical shift dispersion than α85 (data not shown). The amide hydrogen exchange kinetics show a well protected set of about four protons after 20 hours (data not shown). This faster exchange relative to α85 is explained by α85W43V's significantly lower stability

(Mayo & Baldwin, 1993). α85W43V appears to have improved structural specificity at the expense of stability, a phenomenon observed previously in coiled coils (Harbury, *et al.*, 1993). By using an exposure penalty, the design algorithm produced a protein with greater native-like character.

We have quantitatively defined the role of packing specificity in protein design and have provided practical bounds for the role of steric forces in our protein design program. This study differs from previous work because of the use of an objective, quantitative program to vary packing forces during design, which allows us to readily apply our conclusions to different protein systems. Further, by using the minimum effective level of steric forces, we were able to design a wider variety of packing arrangements that were compatible with the given fold. Finally, we have identified a difficulty in the design of side chains that lie at the boundary between the core and the surface of a protein, and we have implemented a nonpolar surface exposure penalty in our sequence design scoring function that addresses this problem.

Example 5

Design of a full protein

The entire amino acid sequence of a protein motif has been computed. As in Example 4, the second zinc finger module of the DNA binding protein Zif268 was selected as the design template. In order to assign the residue positions in the template structure into core, surface or boundary classes, the orientation of the Cα-Cβ vectors was assessed relative to a solvent accessible surface computed using only the template Cα atoms. A solvent accessible surface for only the Cα atoms of the target fold was generated using the Connolly algorithm with a probe radius of 8.0 Å, a dot density of 10 $Å^2$, and a Cα radius of 1.95 Å. A residue was classified as a core position if the distance from its Cα, along its Cα-Cβ vector, to the solvent accessible surface was greater than 5Å, and if the distance from its Cβ to the nearest surface point was greater than 2.0 Å. The remaining residues were classified as surface positions if the sum of the distances from their Cα, along their Cα-Cβ vector, to the solvent accessible surface plus the distance from their Cβ to the nearest surface point was less than 2.7 Å. All remaining residues were classified as boundary positions. The classifications for Zif268 were used as computed except that positions 1, 17 and 23 were converted from the boundary to the surface class to account for end effects from the proximity of chain termini to these residues in the teriary structure and inaccuracies in the assignment.

The small size of this motif limits to one (position 5) the number of residues that can be assigned unambiguously to the core while seven residues (positions 3, 7, 12, 18, 21, 22, and 25) were classified as boundary and the remaining 20 residues were assigned to the surface. Interestingly, while three of the zinc binding positions of Zif268 are in the boundary or core, one residue, position 8, has a Cα-Cβ vector directed away from the protein's geometric center and is classified as a surface position. As in our previous studies, the amino acids considered at the core positions during sequence selection were A, V, L, I, F, Y, and W; the amino acids considered at the surface positions were A, S, T, H, D, N, E, Q, K, and R; and the combined core and surface amino acid sets (16 amino acids) were considered at the boundary positions. Two of the residue positions (9 and 27) have φ

angles greater than 0° and are set to Gly by the sequence selection algorithm to minimize backbone strain.

The total number of amino acid sequences that must be considered by the design algorithm is the product of the number of possible amino acid types at each residue position. The ββα motif residue classification described above results in a virtual combinatorial library of 1.9×10^{27} possible amino acid sequences (one core position with 7 possible amino acids, 7 boundary positions with 16 possible amino acids, 18 surface positions with 10 possible amino acids and 2 positions with φ angles greater than 0° each with 1 possible amino acid). A corresponding peptide library consisting of only a single molecule for each 28 residue sequence would have a mass of 11.6 metric tons. In order to accurately model the geometric specificity of side-chain placement, we explicitly consider the torsional flexibility of amino acid side chains in our sequence scoring by representing each amino acid with a discrete set of allowed conformations, called rotamers. As above, a backbone dependent rotatmer library was used (Dunbrack and Karplus, supra), with adjustments in the χ_1 and χ_2 angles of hydrophobic residues. As a result, the design algorithm must consider all rotamers for each possible amino acid at each residue position. The total size of the search space for the ββα motif is therefore 1.1×10^{62} possible rotamer sequences. The rotamer optimization problem for the ββα motif required 90 CPU hours to find the optimal sequence.

The optimal sequence, shown in Figure11, is called Full Sequence Design-1 (FSD-1) (SEQ ID NO:3). Even though all of the hydrophilic amino acids were considered at each of the boundary positions, the algorithm selected only nonpolar amino acids. The eight core and boundary positions are predicted to form a well-packed buried cluster. The Phe side chains selected by the algorithm at the zinc binding His positions of Zif268, positions 21 and 25, are over 80% buried and the Ala at position 5 is 100% buried while the Lys at position 8 is greater than 60% exposed to solvent. The other boundary positions demonstrate the strong steric constraints on buried residues by packing similar side chains in an arrangement similar to that of Zif268. The calculated optimal configuration for core and boundary residues buries ~1150 $Å^2$ of nonpolar surface area. On the helix surface, the program positions Asn 14 as a helix N-cap with a hydrogen bond between its side-chain carbonyl oxygen and the backbone amide proton of residue 16. The eight charged residues on the helix form three pairs of hydrogen bonds, though in our coiled coil designs helical surface hydrogen bonds appeared to be less important than the overall helix propensity of the sequence (Dahiyat, *et al.*, Science (1997)). Positions 4 and 11 on the exposed sheet surface were selected to be Thr, one of the best β-sheet forming residues (Kim, *et al.* 1993).

Figure 11 shows the alignment of the sequences for FSD-1 (SEQ ID NO:3) and Zif268 (SEQ ID NO:1). Only 6 of the 28 residues (21%) are identical and only 11 (39%) are similar. Four of the identities are in the buried cluster, which is consistent with the expectation that buried residues are more conserved than solvent exposed residues for a given motif (Bowie, *et al.*, Science **247**:1306-1310 (1990)). A BLAST (Altschul, *et al.*, supra) search of the FSD-1 sequence (SEQ ID NO:3) against the non-redundant protein sequence database of the National Center for Biotechnology

Information did not find any zinc finger protein sequences. Further, the BLAST search found only low identity matches of weak statistical significance to fragments of various unrelated proteins. The highest identity matches were 10 residues (36%) with p values ranging from 0.63 - 1.0. Random 28 residue sequences that consist of amino acids allowed in theββα position classification described above produced similar BLAST search results, with 10 or 11 residue identities (36 - 39%) and p values ranging from 0.35 - 1.0, further suggesting that the matches found for FSD-1 are statistically insignificant. The very low identity to any known protein sequence demonstrates the novelty of the FSD-1 sequence (SEQ ID NO:3) and underscores that no sequence information from any protein motif was used in our sequence scoring function.

In order to examine the robustness of the computed sequence, the sequence of FSD-1 (SEQ ID NO:3) was used as the starting point of a Monte Carlo simulated annealing run. The Monte Carlo search finds high scoring, suboptimal sequences in the neighborhood of the optimal solution (Dahiyat, *et al.*, (1996) (supra)). The energy spread from the ground-state solution to the 1000^{th} most stable sequence is about 5 kcal/mol indicating that the density of states is high. The amino acids comprising the core of the molecule, with the exception of position 7, are essentially invariant (Figure 11). Almost all of the sequence variation occurs at surface positions, and typically involves conservative changes. Asn 14, which is predicted to form a helix N-cap, is among the most conserved surface positions. The strong sequence conservation observed for critical areas of the molecule suggests that if a representative sequence folds into the design target structure, then perhaps thousands of sequences whose variations do not disrupt the critical interactions may be equally competent. Even if billions of sequences would successfully achieve the target fold, they would represent only a vanishingly small proportion of the 10^{27} possible sequences.

Experimental validation. FSD-1 was synthesized in order to characterize its structure and assess the performance of the design algorithm. The far UV circular dichroism (CD) spectrum of FSD-1 shows minima at 220 nm and 207 nm, which is indicative of a folded structure (data not shown). The thermal melt is weakly cooperative, with an inflection point at 39 °C, and is completely reversible (data not shown). The broad melt is consistent with a low enthalpy of folding which is expected for a motif with a small hydrophobic core. This behavior contrasts the uncooperative thermal unfolding transitions observed for other folded short peptides (Scholtz, *et al.*, 1991). FSD-1 is highly soluble (greater than 3 mM) and equilibrium sedimentation studies at 100 μM, 500 μM and 1 mM show the protein to be monomeric. The sedimentation data fit well to a single species, monomer model with a molecular mass of 3630 at 1 mM, in good agreement with the calculated monomer mass of 3488. Also, far UV CD spectra showed no concentration dependence from 50 μM to 2 mM, and nuclear magnetic resonance (NMR) COSY spectra taken at 100 μM and 2 mM were essentially identical.

The solution structure of FSD-1 was solved using homonuclear 2D ^{1}H NMR spectroscopy (Piantini, *et al.*, 1982). NMR spectra were well dispersed indicating an ordered protein structure and easing resonance assignments. Proton chemical shift assignments were determined with standard

homonuclear methods (Wuthrich, 1986). Unambiguous sequential and short-range NOEs indicate helical secondary structure from residues 15 to 26 in agreement with the design target.

The structure of FSD-1 was determined using 284 experimental restraints (10.1 restraints per residue) that were non-redundant with covalent structure including 274 NOE distance restraints and 10 hydrogen bond restraints involving slowly exchanging amide protons. Structure calculations were performed using X-PLOR (Brunger, 1992) with standard protocols for hybrid distance geometry-simulated annealing (Nilges, *et al.*, FEBS Lett. 229:317 (1988)). An ensemble of 41 structures converged with good covalent geometry and no distance restraint violations greater than 0.3 Å (Table 9).

Table 9. NMR structure determination: distance restraints, structural statistics and atomic root-mean-square (rms) deviations. <SA> are the 41 simulated annealing structures, SA is the average structure before energy minimization, $(SA)_r$ is the restrained energy minimized average structure, and SD is the standard deviation.

Distance restraints		
Intraresidue	97	
Sequential	83	
Short range (□i-j□ = 2-5 residues)	59	
Long range (□i-j□ > 5 residues)	35	
Hydrogen bond	10	
Total	284	
Structural statistics		
	<SA> ± SD	$(SA)_r$
Rms deviation from distance restraints (Å)	0.043 ± 0.003	0.038
Rms deviation from idealized geometry		
Bonds (Å)	0.0041 ± 0.0002	0.0037
Angles (degrees)	0.67 ± 0.02	0.65
Impropers (degrees)	0.53 ± 0.05	0.51
Atomic rms deviations (Å)*		
	<SA> vs. SA ± SD	<SA> vs. $(SA)_r$ ± SD
Backbone	0.54 ± 0.15	0.69 ± 0.16
Backbone + nonpolar side chains†	0.99 ± 0.17	1.16 ± 0.18
Heavy atoms	1.43 ± 0.20	1.90 ± 0.29

*Atomic rms deviations are for residues 3 to 26, inclusive. Residues 1, 2, 27 and 28 were disordered (φ, ψ angular order parameters (34) < 0.78) and had only sequential and i-j = 2 NOEs. †Nonpolar side chains are from residues 3, 5, 7, 12, 18, 21, 22, and 25 which consitute the core of the protein.

The backbone of FSD-1 is well defined with a root-mean-square (rms) deviation from the mean of 0.54 Å (residues 3-26). Considering the buried side chains (residues 3, 5, 7, 12, 18, 21, 22, and 25) in addition to the backbone gives an rms deviation of 0.99 Å, indicating that the core of the molecule is well ordered. The stereochemical quality of the ensemble of structures was examined using PROCHECK (Laskowski, *et al.*, J. Appl. Crystallogr. 26:283 (1993)). Not including the disordered termini and the glycine residues, 87% of the residues fall in the most favored region and the remainder in the allowed region of φ, ψ space. Modest heterogeneity is present in the first strand (residues 3-6) which has an average backbone angular order parameter (Hyberts, *et al.*, 1992) of <S> = 0.96 ± 0.04 compared to the second strand (residues 9-12) with an <S> = 0.98 ± 0.02 and the helix (residues 15-26) with an <S> = 0.99 ± 0.01. Overall, FSD-1 is notably well ordered and, to our knowledge, is the shortest sequence consisting entirely of naturally occurring amino acids that folds to a unique structure without metal binding, oligomerization or disulfide bond formation (McKnight, *et al.*, 1997).

The packing pattern of the hydrophobic core of the NMR structure ensemble of FSD-1 (Tyr 3, Ile 7, Phe 12, Leu 18, Phe 21, Ile 22, and Phe 25) is similar to the computed packing arrangement. Five of the seven residues have χ_1 angles in the same gauche$^+$, gauche$^-$ or trans category as the design target, and three residues match both χ_1 and χ_2 angles. The two residues that do not match their computed χ_1 angles are Ile 7 and Phe 25, which is consistent with their location at the less constrained, open end of the molecule. Ala 5 is not involved in its expected extensive packing interactions and instead exposes about 45% of its surface area because of the displacement of the strand 1 backbone relative to the design template. Conversely, Lys 8 behaves as predicted by the algorithm with its solvent exposure (60%) and χ_1 and χ_2 angles matching the computed structure. Most of the solvent exposed residues are disordered which precludes examination of the predicted surface residue hydrogen bonds. Asn 14, however, forms a helix N-cap from its sidechain carbonyl oxygen as predicted, but to the amide of Glu 17, not Lys 16 as expected from the design. This hydrogen bond is present in 95% of the structure ensemble and has a donor-acceptor distance of 2.6 ± 0.06 Å. In general, the side chains of FSD-1 correspond well with the design program predictions.

A comparison of the average restrained minimized structure of FSD-1 and the design target was done (data not shown). The overall backbone rms deviation of FSD-1 from the design target is 1.98 Å for residues 3-26 and only 0.98 Å for residues 8-26 (Table 10).

Table 10. Comparison of the FSD-1 experimentally determined structure and the design target structure. The FSD-1 structure is the restrained energy minimized average from the NMR structure determination. The design target structure is the second DNA binding module of the zinc finger Zif268 (9).

	Atomic rms deviations (Å)
Backbone, residues 3-26	1.98

Atomic rms deviations (Å)		
Backbone, residues 8-26	0.98	
Super-secondary structure parameters*		
	FSD-1	Design Target
h (Å)	9.9	8.9
θ (degrees)	14.2	16.5
Ω (degrees)	13.1	13.5

*h, θ, Ω are calculated as previously described (36, 37). *h* is the distance between the centroid of the helix Cα coordinates (residues 15-26) and the least-square plane fit to the Cα coordinates of the sheet (residues 3-12. *θ* is the angle of inclination of the principal moment of the helix Cα atoms with the plane of the sheet. *Ω* is the angle between the projection of the principal moment of the helix onto the sheet and the projection of the average least-square fit line to the strand Cα coordinates (residues 3-6 and 9-12) onto the sheet.

The largest difference between FSD-1 and the target structure occurs from residues 4-7, with a displacement of 3.0-3.5 Å of the backbone atom positions of strand 1. The agreement for strand 2, the strand to helix turn, and the helix is remarkable, with the differences nearly within the accuracy of the structure determination. For this region of the structure, the rms difference of φ,ψ angles between FSD-1 and the design target is only 14 ± 9°. In order to quantitatively assess the similarity of FSD-1 to the global fold of the target, we calculated their supersecondary structure parameters (Table 9) (Janin & Chothia, J. Mol. Biol. 143:95 (1980); Su & Mayo, Protein Sci. in press, 1997), which describe the relative orientations of secondary structure units in proteins. The values of θ, the inclination of the helix relative to the sheet, and Ω, the dihedral angle between the helix axis and the strand axes, are nearly identical. The height of the helix above the sheet, h, is only 1 Å greater in FSD-1. A study of protein core design as a function of helix height for Gβ1 variants demonstrated that up to 1.5 Å variation in helix height has little effect on sequence selection (Su & Mayo, supra, 1997). The comparison of secondary structure parameter values and backbone coordinates highlights the excellent agreement between the experimentally determined structure of FSD-1 and the design target, and demonstrates the success of our algorithm at computing a sequence for this ββα motif.

The quality of the match between FSD-1 and the design target demonstrates the ability of our program to design a sequence for a fold that contains the three major secondary structure elements of proteins: sheet, helix, and turn. Since the ββα fold is different from those used to develop the sequence selection methodology, the design of FSD-1 represents a successful transfer of our program to a new motif.

Example 6

Calculation of solvent accessible surface area scaling factors

In contrast to the previous work, backbone atoms are included in the calculation of surface areas. Thus, the calculation of the scaling factors proceeds as follows.

The program BIOGRAF (Molecular Simulations Incorporated, San Diego, California) was used to generate explicit hydrogens on the structures which were then conjugate gradient minimized for 50 steps using the DREIDING force field. Surface areas were calculated using the Connolly algorithm with a dot density of 10 Å-2, using a probe radius of zero and an add-on radius of 1.4 Å and atomic radii from the DREIDING force-field. Atoms that contribute to the hydrophobic surface area are carbon, sulfur and hydrogen atoms attached to carbon and sulfur.

For each side-chain rotamer *r* at residue position *i* with a local tri-peptide backbone *t*3, we calculated $A^{0}_{i,t3}$ the exposed area of the rotamer and its backbone in the presence of the local tri-peptide backbone, and $A_{i,t}$ the exposed area of the rotamer and its backbone in the presence of the entire template *t* which includes the protein backbone and any side-chains not involved in the calculation (Figure 13). The difference between $A^{0}_{i,t3}$, and $A_{i,t}$ is the total area buried by the template for a rotamer *r* at residue position *i*. For each pair of residue positions *i* and *j* and rotamers *r* and *s* on *i* and *j*, respectively, $A_{i,j,t}$ the exposed area of the rotamer pair in the presence of the entire template, is calculated. The difference between $A_{i,j,t}$ and the sum of $A_{i,t}$ and $A_{j,t}$ is the area buried between residues *i* and *j*, excluding that area by the template. The pairwise approximation to the total buried surface area is:

Equation 29:

$$A^{pairwise}_{buried} = \sum_{i} (A^{0}_{i,t3} - A_{i,t}) + f \sum_{i<j} (A_{i,t} + A_{j,t} - A_{i,j,t})$$

As shown in Figure 13, the second sum in Equation 29 over-counts the buried area. We have therefore multiplied the second sum by a scale factor f whose value is to be determined empirically. Expected values of f are discussed below.

Noting that the buried and exposed areas should add to the total area, $\Sigma_i A^{0}_{i,t3}$, the solvent-exposed surface area is:

Equation 30:

$$A^{pairwise}_{exposed} = \sum_{i} A_{i,t} - f \sum_{i<j} (A_{i,t} + A_{j,t} - A_{i,j,t})$$

The first sum of Equation 30 represents the total exposed area of each rotamer in the context of the protein template ignoring interactions with other rotamers. The second sum of Equation 30 subtracts the buried areas between rotamers and is scaled by the same parameter f as in Equation 29.

Some insight into the expected value of f can be gained from consideration of a close-packed face centered cubic lattice of spheres or radius *r*. When the radii are increased from *r* to *R*, the surface

area on one sphere buried by a neighboring sphere is $2\pi R(R - r)$. We take r to be a carbon radius (1.95 Å), and R is 1.4 Å larger. Then, using:

$$f = \frac{true\ buried\ area}{pairwise\ buried\ area}$$

and noting that each sphere has 12 neighbors, results in:

$$f = \frac{4\pi R_2}{12x2\pi R(R - r)}$$

This yields f = 0.40. A close-packed face centered cubic lattice has a packing fraction of 74%. Protein interiors have a similar packing fraction, although because many atoms are covalently bonded the close packing is exaggerated. Therefore this value of f should be a lower bound for real protein cores. For non-core residues, where the packing fraction is lower, a somewhat larger value of f is expected.

We classified residues from ten proteins ranging in size from 54 to 289 residues into core or non-core as follows. We classified resides as core or non-core using an algorithm that considered the direction of each side-chain's Cα-Cβ vector relative to the surface computed using only the template Cα atoms with a carbon radius of 1.95 Å, a probe radius of 8 Å and no add-on radius. A residue was classified as a core position if both the distance from its Cα atom (along its Cα-Cβ vector) to the surface was greater than 5.0 Å and the distance from its Cβ atom to the nearest point on the surface was greater than 2.0 Å. The advantage of such an algorithm is that a knowledge of the amino acid type actually present at each residue position is not necessary. The proteins were as shown in Table I, showing selected proteins, total number of residues and the number of residues in the core and non-core of each protein (Gly and pro were not considered).

Brookhaven Identifier	Total Size	Core Size	Non-Core Size
1enh	54	10	40
1pga	56	10	40
1ubi	76	16	50
1mol	94	19	61
1kpt	105	27	60
4azu-A	128	39	71
1gpr	158	39	89
1gcs	174	53	98
1edt	266	95	133
1pbn	289	96	143

The classification into core and non-core was made because core residues interact more strongly with one another than do non-core residues. This leads to greater over-counting of the buried surface area for core residues.

Considering the core and non-core cases separately, the value of f which most closely reproduced the true Lee and Richards surface areas was calculated for the ten proteins. The pairwise approximation very closely matches the true buried surface area (data not shown). It also performs very well for the exposed hydrophobic surface area of non-core residues (data not shown). The calculation of the exposed surface area of the entire core of a protein involves the difference of two large and nearly equal areas and is less accurate; as will be shown, however, when there is a mixture of core and non-core residues, a high accuracy can still be achieved. These calculations indicate that for core residues f is 0.42 and for non-core residues f is 0.79.

To test whether the classification of residues into core and non-core was sufficient, we examined subsets of interacting residues in the core and non-core positions, and compared the true buried area of each subset with that calculated (using the above values of f). For both subsets of the core and the non-core, the correlation remained high (R^2 = 1.00) indicating that no further classification is necessary (data not shown). (Subsets were generated as follows: given a seed residue, a subset of size two was generated by adding the closest residue: the next closest residue was added for a subset of size three, and this was repeated up to the size of the protein. Additional subsets were generated by selecting different seed residues.)

It remains to apply this approach to calculating the buried or exposed surface areas of an arbitrary selection of interacting core and non-core residues in a protein. When a core residue and a non-core residue interact, we replace Equation 29 with:

Equation 31:

$$A_{buried}^{pairwise} = \sum_i (A^0_{i,t3} - A_{i,t}) + \sum_{i<j} (f_i A_{i,t} + f_j A_{j,t} - f_{ij} A_{i,j,t})$$

and Equation 30 with Equation 32:

$$A_{exposed}^{pairwise} = \sum_i A_{i,t} - \sum_{i<j} (f_i A_{i,t} + f_j A_{j,t} - f_{ij} A_{i,j,t})$$

where f_i and f_j are the values of f appropriate for residues *i* and *j*, respectively, and $f_{\{ij\}}$ takes on an intermediate value. Using subsets from the whole of 1pga, the optimal value of f_{ij} was found to be 0.74. This value was then shown to be appropriate for other test proteins (data not shown).

Example 7

The use of supersecondary structure parameters to incorporate backbone flexibility

This example is concerned primarily with coupling backbone flexibility and the selection of amino acids for protein cores and an assessment of the tolerance of our side-chain selection algorithm to perturbations in protein backbone geometry. An ideal model system for these purposes is the β1 immunoglobulin-binding domain of streptococcal protein G (Gβ1) (Gronenborn et al., *Science 253*:657–661(1991 "A novel, highly stable fold of the immunoglobulin binding domain of streptococcal

protein G"). Its small size, 56 residues, renders computations more tractable and simplifies production of the protein by either synthetic or recombinant methods. A solution structure (Gronenborn et al., *id*) and several crystal structures (Gallagher et al., *Biochemistry 33:*4721–4729 (1994), "Two crystal structures of the β1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR") are available to provide backbone templates for the side-chain selection algorithm. In addition, the energetics and structural dynamics of Gβ1 have been extensively characterized (Alexander et al. *Biochemistry 31:*3597–3603, (1992) "Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains β1 and β2 — Why small proteins tend to have high denaturation temperatures"); Barchi et al., *Protein Sci 3:*15–21 (1994)"Investigation of the backbone dynamics of the IgG-binding domain of spreptococcal protein G by heteronuclear two-dimensional ^{1}H-^{15}N nuclear magnetic resonance spectroscopy"); Kuszewski et al., *Protein Sci 3:*1945–1952 (1994) "Fast folding of a prototypic polypeptide — The immunoglobulin binding domain of streptococcal protein G"); Orban et al., *Biochemistry 34:*15291–15300 (1995) "Assessment of stability differences in the protein G β1 and β2 domains from hydrogen deuterium exchange — Comparison with calorimetric data"). Gβ1 contains no disulfide bonds and does not require a cofactor or metal ion to fold, but relies upon the burial of its hydrophobic core for stability. Further, Gβ1 contains sheet, helix and turn structures and is without the repetitive side-chain packing patterns found in coiled coils and some helical bundles. This lack of periodicity reduces the bias from a particular secondary or tertiary structure and necessitates the use of an objective algorithm for side-chain selection. Perhaps most important for this study, the Gβ1 backbone can be classified as an α/β fold, a class for which extensive super-secondary structure analysis has been performed (Chothia et al., 1977 (*id*); Janin & Chothia, 1980 (*id*); Cohen et al., 1982 (*id*); Chou et al., 1985 (*id*)).

Sequence positions that constitute the core were chosen by examining the side-chain solvent accessible surface area of Gβ1. We selected the ten most buried positions which includes residues 3, 5, 7, 20, 26, 30, 34, 39, 52 and 54. The remainder of the protein structure, including all other side chains and the backbone, was used as the template for sequence selection calculations at the ten core positions.

Backbone perturbation, scoring functions and DEE

The initial Gβ1 structure was taken from PDB entry 1pga (Bernstein et al., *J. Mol Biol 112:*535–542 (1977) "The Protein Data Bank: A computer-based archival file for macromolecular structures"); Gallagher et al., 1994 (*id*). The program BIOGRAF (Molecular Simulations Incorporated, San Diego, CA) was used to generate explicit hydrogens on the structure which was then conjugate gradient minimized for 50 steps using the DREIDING forcefield (Mayo et al., 1990 (*id*)). The coordinate positions of atoms not involved in core sequence selection or backbone perturbations were kept fixed. Concerted backbone movements were performed by repositioning the α-helix (residues 23 through 36) to reflect the desired change in the indicated super-secondary structure parameter value. The coordinate positions of atoms belonging to residues 23, 24, 25, 27, 28, 29, 31, 32, 33, 35 and 36 were kept fixed after repositioning the helix. The distorted peptide bonds that result from backbone perturbations were left unchanged. The Δh, $\Delta\Omega$ and $\Delta\sigma$-series perturbations were carried out by

translating the helix along the sheet axis, rotating the helix about the sheet axis and rotating the helix about the vector parallel to the helix axis that passes through the helix center, respectively (see Fig. 14). The Δθ-series perturbations were carried out by rotating the helix about the vector resulting from the cross product of the sheet axis and the vector parallel to the helix axis that passes through the helix center. A Lennard-Jones 12-6 potential was used for van der Waals interactions with atomic radii scaled by either 1.0 or 0.9 as indicated (Dahiyat & Mayo, submitted). The Richards definition of solvent-accessible surface area (Lee & Richards, 1971, supra) was used and areas were calculated with the Connolly algorithm (Connolly, 1983, supra). An atomic solvation parameter of 23.2 cal/mol/Å^2 was used to favor hydrophobic burial (Dahiyat & Mayo, 1996, supra). The rotamer library and DEE optimization followed the methods of our previous work (Dahiyat & Mayo, 1996, supra). Calculations were performed on either a 12 processor, R10000-based Silicon Graphics Power Challenge or a 512 node Intel Delta.

Mutagenesis and protein purification

A synthetic Gβ1 gene (Minor & Kim, 1994) was cloned into pET11a (Novagen) and used as the template for inverse PCR mutagenesis (Hemsley et al., 1989). 5' phosphorylated oligos (Genosys) were used without further purification. Mutant sequences were confirmed by DNA sequencing. The expression and purification of the mutant proteins followed published procedures (Minor & Kim, 1994). Incomplete N-terminal processing resulted in a mixture of 56 and 57 residue proteins which were separated by HPLC (Minor & Kim, 1994, supra). The 57 residue proteins were used in all cases except for mutants $\Delta h_{0.9}$[-1.50Å] and $\Delta h_{0.9}$[+1.50Å], where the 56 residue proteins were used. Molecular weights were confirmed by mass spectrometry.

CD and NMR

CD spectra were measured on an Aviv 62DS spectrometer at pH 6.0, 50 mM sodium phosphate buffer, 25 °C and 50 μM protein. A 1 mm pathlength cell was used and the temperature was controlled by a thermoelectric unit. Thermal melts were performed in the same buffer using two degree temperature increments with an averaging time of 10 s and an equilibration time of 90 s. T_m values were derived from the ellipticity at 218 nm ($[\theta]_{218}$) by evaluating the maximum of a $d[\theta]_{218}/dT$ versus T plot. The T_m's were reproducible to within two degrees. Protein concentrations were determined by UV spectrophotometry. NMR samples were prepared in 90/10 H_2O/D_2O and 50 mM phosphate buffer at pH 6.0. Spectra were acquired on a Varian Unity Plus 600 MHz spectrometer at 25 °C. 1024 transients were acquired with 1.5 seconds of solvent presaturation used for water suppression. Samples were approximately 0.5 mM.

Results

Four sets of perturbed backbones were generated by varying Gβ1's super-secondary structure parameter values (Fig. 14). All possible core sequences consisting of alanine, valine, leucine, isoleucine, phenylalanine, tyrosine and tryptophan (A, V, L, I, F, Y and W) were considered for each perturbed backbone. The rotamer library was as described above (see Dahiyat & Mayo, 1996,

supra). Optimizing the sequences of the cores of Gβ1 and its structural homologues with 217 possible hydrophobic rotamers considered at each of the ten core positions results in 217^{10} (~10^{23}) rotamer sequences. Our scoring function consisted of two components: a van der Waals energy term and an atomic solvation term favoring burial of hydrophobic surface area. The van der Waals radii of the atoms in the simulation were scaled by either 1.0 or 0.9 in order to reduce the effects of using discrete rotamers (see Mayo et al., 1990, supra, and Example 6). Global optimum sequences for each of the backbone variants were found using the Dead-End Elimination (DEE) theorem (Desmet et al., 1992, supra; Desmet et al., 1994, supra; Goldstein, 1994, supra). Optimal sequences, and their corresponding proteins, are named by the backbone perturbation type, the size of the perturbation and the radius scale factor used in their design. For example, the sequence designed using a template whose helix was translated by +1.50 Å along the sheet axis and a radius scale factor of 0.9 is called $\Delta h_{0.9}$[+1.50Å]. Backbone perturbations that result in the same calculated core sequence are named by the perturbation with the greatest magnitude. For example, $\Delta h_{0.9}$ backbone perturbations of +1.25 and +1.50 Å result in the same sequence which is called $\Delta h_{0.9}$[+1.50Å]. The calculated core sequences corresponding to various backbone perturbations are listed in Tables 1-5, below.

Table 11. *DEE determined optimal sequences for the core positions of Gβ1 as a function of $\Delta h_{0.9}$*[a]

		Gβ1 sequence (SEQ ID NO:38)											
$\Delta h_{0.9}$ (Å)	vol	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	PHE	VAL	Tm (°C)	NMR
		3	5	7	20	26	30	34	39	52	54		
-1.50	1.04	PHE	ILE	VAL	VAL	|	|	|	ILE	|	| (SEQ ID NO:80)	69	+
-1.25	1.04	PHE	ILE	VAL	VAL	|	|	|	ILE	|	| (SEQ ID NO:80)	69	+
-1.00	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:81)	89	+
-0.75	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:81)	89	+
-0.50	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:81)	89	+
-0.25	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:81)	89	+
0.00	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:82)	91	+
-0.25	1.05	PHE	|	ILE	|	|	|	|	ILE	TRP	| (SEQ ID NO:83)	89	+
+0.50	1.05	PHE	|	ILE	|	|	|	|	ILE	TRP	| (SEQ ID NO:83)	89	+
+0.75	1.05	PHE	|	ILE	|	|	|	|	ILE	TRP	| (SEQ ID NO:83)	89	+
+1.00	1.13	PHE	|	ILE	|	|	|	ILE	ILE	TRP	| (SEQ ID NO:84)	85	+
+1.25	1.20	PHE	|	ILE	|	LEU	|	ILE	ILE	TRP	| (SEQ ID NO:85)	53	-
+1.50	1.20	PHE	|	ILE	|	LEU	|	ILE	ILE	TRP	| (SEQ ID NO:85)	53	-

[a]The Gβ1 wild-type sequence (SEQ ID NO:38) and position numbers are shown at the top of the Table. A vertical bar indicates identity with the Gβ1 sequence (SEQ ID NO:38). Δh is the change in the super-secondary structure parameter, h; vol is the fraction of core side-chain volume relative to

the Gβ1 sequence (SEQ ID NO:38); T_m is the melting temperature measured by circular dichroism; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra. The T_m's for $\Delta h_{0.9}$[-1.50Å] and $\Delta h_{0.9}$[+1.50Å] were determined for 56 residue proteins (compared to 57 residue proteins for Gβ1 and all other mutants) which overstates the melting temperature by about 2 °C, the melting temperature difference between the 56 and 57 residue versions of Gβ1.

Table 12. DEE determined optimal sequences for the core positions of Gβ1 as a function of Δh_{10}[a]

		Gβ1 sequence (SEQ ID NO:38)											
Δh_{10} (Å)	vol	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	PHE	VAL	Tm (°C)	NMR
		3	5	7	20	26	30	34	39	52	54		
-1.50	0.52	ALA	ALA	ALA	|	|	ALA	|	LEU	ALA	ALA (SEQ ID NO:86)	ND	ND
-1.25	0.62	PHE	ALA	ALA	|	|	ALA	|	LEU	ALA	ALA (SEQ ID NO:87)	ND	ND
-1.00	0.62	PHE	ALA	ALA	|	|	ALA	|	LEU	ALA	ALA(SEQ ID NO:87)	ND	ND
-0.75	0.91	PHE	ALA	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:88)	ND	ND
-0.50	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| ((SEQ ID NO:89)	89	+
-0.25	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:89)	89	+
0.00	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
+0.25	1.05	PHE	|	ILE	|	|	|	|	ILE	TRP	| (SEQ ID NO:91)	89	+
+0.50	1.05	PHE	|	ILE	|	|	|	|	ILE	TRP	| (SEQ ID NO:91)	89	+
+0.75	1.05	PHE	|	ILE	|	|	|	|	ILE	TRP	| (SEQ ID NO:91)	89	+
+1.00	1.05	PHE	|	ILE	|	|	|	|	ILE	TRP	| (SEQ ID NO:91)	89	+
+1.25	1.05	PHE	|	ILE	|	|	|	|	ILE	TRP	| (SEQ ID NO:91)	89	+
+1.50	1.11	PHE	|	ILE	|	|	LEU	ILE	ILE	TRP	| (SEQ ID NO:92)	73	+

[a]The Gβ1 wild-type sequence (SEQ ID NO:38) and position numbers are shown at the top of the Table. A vertical bar indicates identity with the Gβ1 sequence (SEQ ID NO:38). Δh is the change in the super-secondary structure parameter, h; vol is the fraction of core side-chain volume relative to the Gβ1 sequence (SEQ ID NO:38); T_m is the melting temperature measured by circular dichroism; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra; ND indicates a property that was not determined.

Table 13. *DEE determined optimal sequences for the core positions of Gβ1 as a function of* $\Delta\Omega_{0.9}$[a]

		Gβ1 sequence (SEQ ID NO:38)											
ΔΩ (°)	vol	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	PHE	VAL	Tm (°C)	NMR
		3	5	7	20	26	30	34	39	52	54		
-10.0	1.00	VAL	|	VAL	VAL	|	|	|	ILE	|	| (SEQ ID NO:93)	ND	ND
-7.5	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:89)	89	+
-5.0	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:89)	89	+
-2.5	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:89)	89	+
0.0	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
+2.5	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
+5.0	1.06	PHE	|	ILE	VAL	|	|	|	ILE	|	| (SEQ ID NO:94)	ND	ND
+7.5	1.06	PHE	|	ILE	VAL	|	|	|	ILE	|	| (SEQ ID NO:94)	ND	ND
+10.0	1.06	PHE	|	ILE	VAL	|	|	|	ILE	|	| (SEQ ID NO:94)	ND	ND

[a]The Gβ1 wild-type sequence (SEQ ID NO:38) and position numbers are shown at the top of the Table. A vertical bar indicates identity with the Gβ1 sequence (SEQ ID NO:38). ΔΩ is the change in the super-secondary structure parameter, Ω; vol is the fraction of core side-chain volume relative to the Gβ1 sequence (SEQ ID NO:38); T_m is the melting temperature measured by circular dichroism; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra; ND indicates a property that was not determined.

Table 14. DEE determined optimal sequences for the core positions of Gβ1 as a function of $\Delta h_{0.9}$[a]

		Gβ1 sequence (SEQ ID NO:38)											
Δθ_(°)	vol	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	PHE	VAL	Tm (°C)	NMR
		3	5	7	20	26	30	34	39	52	54		
-10.0	0.98	PHE	|	ALA	|	|	|	|	LEU	TRP	| (SEQ ID NO:95)	ND	ND
-7.5	1.00	PHE	|	LEU	|	|	|	|	LEU	TRP	ALA (SEQ ID NO:96)	ND	ND
-5.0	1.03	PHE	|	VAL	|	|	|	|	ILE	TRP	| (SEQ ID NO:97)	ND*	ND*
-2.5	1.03	PHE	|	VAL	|	|	|	|	ILE	TRP	| (SEQ ID NO:97)	ND*	ND*
0.0	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
+2.5	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
+5.0	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
+7.5	1.08	PHE	|	ILE	VAL	|	TRP	|	ILE	LEU	| (SEQ ID NO:98)	ND	ND
+10.0	1.08	PHE	|	ILE	VAL	|	TRP	|	ILE	LEU	| (SEQ ID NO:98)	ND	ND

[a]The Gβ1 wild-type sequence (SEQ ID NO:38) and position numbers are shown at the top of the Table. A vertical bar indicates identity with the Gβ1 sequence (SEQ ID NO:38). Δθ is the change in the super-secondary structure parameter, θ; vol is the fraction of core side-chain volume relative to the Gβ1 sequence (SEQ ID NO:38); T_m is the melting temperature measured by circular dichroism; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra; ND indicates a property that was not determined; ND* indicates a property that was not determined, but that is expected to be "positive" based on sequence similarity to other DEE solutions (see $\Delta h_{0.9}$[+0.75Å]).

Table 15. DEE determined optimal sequences for the core positions of Gβ1 as a function of $\Delta\sigma_{0.9}$[a]

		Gβ1 sequence (SEQ ID NO:38)											
Δσ_(°)	vol	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	PHE	VAL	Tm (°C)	NMR
		3	5	7	20	26	30	34	39	52	54		
-10.0	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
-7.5	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
-5.0	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
-2.5	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
0.0	1.01	PHE	|	ILE	|	|	|	|	ILE	|	| (SEQ ID NO:90)	91	+
+2.5	0.99	PHE	|	VAL	|	|	|	|	ILE	|	| (SEQ ID NO:89)	89	+
+5.0	1.03	PHE	|	VAL	|	|	|	|	ILE	TRP	| (SEQ ID NO:97)	ND*	ND*
+7.5	1.04	PHE	|	VAL	|	|	TYR	|	ILE	TRP	| (SEQ ID NO:99)	ND	ND
+10.0	1.04	PHE	|	VAL	|	|	TYR	|	ILE	TRP	| (SEQ ID NO:99)	ND	ND

[a]The Gβ1 wild-type sequence (SEQ ID NO:38) and position numbers are shown at the top of the Table. A vertical bar indicates identity with the Gβ1 sequence (SEQ ID NO:38). Δσ is the change in the super-secondary structure parameter, σ; vol is the fraction of core side-chain volume relative to the Gβ1 sequence (SEQ ID NO:38); T_m is the melting temperature measured by circular dichroism; NMR is a qualitative indication of the degree of chemical shift dispersion in the 1D ^{1}H NMR spectra; ND indicates a property that was not determined; ND$^+$ indicates a property that was not determined, but that is expected to be "positive" based on sequence similarity to other DEE solutions (see $\Delta h_{0.9}$[+0.75Å]).

The optimal sequence for the ten core positions of Gβ1 (SEQ ID NO:38) that is calculated using the native backbone (i.e., no perturbation) contains three conservative mutations relative to the wild-type sequence (Table 11). Y3F and V39I are likely the result of the hydrophobic surface area burial term in the scoring function. L7I reflects a bias in the rotamer library used for these calculations. The crystal structure of Gβ1 has the leucine at position 7 with a nearly eclipsed χ_2 of 111°. This strained χ_2 is unlikely to be an artifact of the structure determination since it is present in two crystal forms and a solution structure (Gronenborn et al., 1991; Gallagher et al., 1994). Our rotamer library does not contain eclipsed rotamers and no staggered leucine rotamers pack well at this position. Instead, the side-chain selection algorithm chose an isoleucine rotamer that conserves the χ_1 dihedral and is able to pack well. We expect the removal of the strained leucine rotamer to stabilize the protein, a prediction that is tested in the experimental section of this work. The sequences that result from varying individual super-secondary structure parameter values show two notable trends. Small variations in the parameter values tend to have little or no effect on the calculated sequences. For example, varying $\Delta h_{0.9}$ from -0.25 to -1.00 Å (Table 4 11) and $\Delta h_{1.0}$ from +0.25 to +1.25 Å (Table 2) has no effect on the calculated sequences which demonstrates the side-chain selection algorithm's tolerance to small variations in the initial backbone geometry. Large variations in the parameter

values tend to result in greater sequence diversity. For example, $\Delta h_{1.0}$[+1.50Å] contains six out of ten possible mutations relative to Gβ1 (Table 12). The apparently anomalous result that occurs for $\Delta h_{0.9}$ at -1.25 and -1.50 Å, an increase in core volume, is explained by the observation that translating the helix towards the sheet plane results in creating a pocket of space in the vicinity of position 20 that ultimately leads to the observed A20V mutation.

Experimental validation of the designed cores focused on seven of the Δh-series mutants which contain between three and six sequence changes relative to Gβ1. The designed sequences resulting from ΔΩ, Δθ and Δσ perturbations are, however, in many cases identical to various Δh-series sequences. Typical far UV circular dichroism (CD) spectra are shown in Figure 15. $\Delta h_{0.9}$[-1.00Å], $\Delta h_{0.9}$[0.00Å], $\Delta h_{0.9}$[+0.75Å] and $\Delta h_{0.9}$[+1.00Å] have CD spectra that are indistinguishable from that of Gβ1 while $\Delta h_{0.9}$[+1.50Å], $\Delta h_{1.0}$[+1.50Å] and $\Delta h_{0.9}$[-1.50Å] have CD spectra similar to that of Gβ1 suggesting that all of the mutants have a secondary structure content similar to the wild-type protein. Thermal melts monitored by CD are shown in Figure 16. All of the mutants have cooperative transitions with melting temperatures (T_m's) ranging from 53 °C for $\Delta h_{0.9}$[+1.50Å] to 91 °C for $\Delta h_{0.9}$[0.00Å] (Table 11). The T_m for Gβ1 is 85°C. The measured T_m's for $\Delta h_{0.9}$[-1.50Å] and $\Delta h_{0.9}$[+1.50Å] are for 56 residue proteins compared to 57 residue proteins in all other cases (see Methods and materials) which results in T_m's that are estimated to be about 2 °C higher than what would be expected for the corresponding 57 residue proteins based on the T_m difference between the 56 and 57 residue versions of Gβ1. The removal of the strained leucine at position seven (L7I) along with the increased hydrophobic burial generated by the Y3F and V39I mutations in $\Delta h_{0.9}$[0.00Å] result in a protein that is measurable more stable than wild-type Gβ1. The extent of chemical shift dispersion in the 1D ^{1}H NMR spectrum of each mutant was assessed to gauge each protein's degree of native-like character (Fig. 5). All of the mutants, except $\Delta h_{0.9}$[+1.50Å], have NMR spectra with chemical shift dispersion similar to that of Gβ1 suggesting that the proteins form well-ordered structures. $\Delta h_{0.9}$[+1.50Å] has a spectrum with broad peaks and no dispersion, which is indicative of a collapsed but disordered and fluctuating structure or non-specific association. All seven mutant proteins retain their ability to bind IgG as measured by binding to an IgG-Sepharose affinity column. The stability and native-like character of $\Delta h_{0.9}$[-1.50Å] and $\Delta h_{1.0}$[+1.50Å] indicate that the sequence selection algorithm is sufficiently robust to tolerate Δh perturbations that are as large as 15% of Gβ1's native height super-secondary structure parameter value of 10 Å.

Although structures have not yet been determined for the six mutants that show good chemical shift dispersion in their NMR spectra, the magnitude of the backbone perturbations used to calculate these sequences are similar to the backbone perturbations observed for core mutations in other proteins (Baldwin et al., 1993; Lim et al., 1994). Elucidation of the structures of several of the mutants should contribute to our general understanding of the deformation modes available to protein backbones of the α/β class and should help define ranges of super-secondary structure parameter value perturbations that will be useful in future sequence calculations.

SEQUENCE LISTING

CLAIMS

We claim:

1. A method executed by a computer under the control of a program, said computer including a memory for storing said program, said method comprising the steps of:

(A) receiving a protein backbone structure with variable residue positions;

(B) establishing a group of potential rotamers for each of said variable residue positions, wherein at least one variable residue position has rotamers from at least two different amino acid side chains; and

(C) analyzing the interaction of each of said rotamers with all or part of the remainder of said protein backbone structure to generate a set of optimized protein sequences, wherein said analyzing step includes a Dead-End Elimination (DEE) computation.

2. A method executed by a computer under the control of a program, said computer including a memory for storing said program, said method comprising the steps of:

(A) receiving a protein backbone structure with variable residue positions;

(B) classifying each variable residue position as either a core, surface or boundary residue;

(C) establishing a group of potential rotamers for each of said variable residue positions, wherein at least one variable residue position has rotamers from at least two different amino acid side chains; and

(D) analyzing the interaction of each of said rotamers with all or part of the remainder of said protein to generate a set of optimized protein sequences.

3. A method according to claim 2 wherein said analyzing step comprises a DEE computation.

4. A method according to claim 1 or 2 wherein said set of optimized protein sequences comprises the globally optimal protein sequence.

5. A method according to claim 1 or 3 wherein said DEE computation is selected from the group consisting of original DEE and Goldstein DEE.

6. A method according to claim 1 or 2 wherein said analyzing step includes the use of at least one scoring function.

7. A method according to claim 6 wherein said scoring function is selected from the group consisting of a Van der Waals potential scoring function, a hydrogen bond potential scoring function,

an atomic solvation scoring function, an electrostatic scoring function and a secondary structure propensity scoring function.

8. A method according to claim 6 wherein said analyzing step includes the use of at least two scoring functions.

9. A method according to claim 6 wherein said analyzing step includes the use of at least three scoring functions.

10. A method according to claim 6 wherein said analyzing step includes the use of at least four scoring functions.

11. A method according to claim 6 wherein said atomic solvation scoring function includes a scaling factor that compensates for over-counting.

12. A method according to claim 1 or 2 further comprising testing at least one member of said set to produce experimental results.

13. A method according to claim 4 further comprising

(D) generating a rank ordered list of additional optimal sequences from said globally optimal protein sequence.

14. A method according to claim 13 wherein said generating includes the use of a Monte Carlo search.

15. A method according to claim 2 wherein said analyzing step step comprises a Monte Carlo computation.

16. A method according to claim 13 further comprising:

(E) testing some or all of said protein sequences from said ordered list to produce potential energy test results.

17. A method according to claim 16 further comprising:

(F) analyzing the correspondence between said potential energy test results and theoretical potential energy data.

18. A method according to claim 1 or 2 further comprising altering at least one supersecondary structure parameter value of said protein backbone structure prior to establishing said potential rotamer group.

19. An optimized protein sequence generated by the method of claim 1 or 2.

20. A nucleic acid sequence encoding a protein sequence according to claim 19.

21. An expression vector comprising the nucleic acid of claim 20.

22. A host cell comprising the nucleic acid of claim 20.

23. A protein having a sequence that is at least about 5% different from a known protein sequence and is at least 20% more stable than the known protein sequence.

24. A computer readable memory to direct a computer to function in a specified manner, comprising:

a side chain module to correlate a group of potential rotamers for residue positions of a protein backbone model;

a ranking module to analyze the interaction of each of said rotamers with all or part of the remainder of said protein to generate a set of optimized protein sequences.

25. A computer readable memory according to claim 24 wherein said ranking module includes a van der Waals scoring function component.

26. A computer readable memory according to claim 24 wherein said ranking module includes an atomic solvation scoring function component.

27. A computer readable memory according to claim 24 wherein said ranking module includes a hydrogen bond scoring function component.

28. A computer readable memory according to claim 24 wherein said ranking module includes a secondary structure scoring function component.

29. A computer readable memory according to claim 24 further comprising

an assessment module to assess the correspondence between potential energy test results and theoretical potential energy data.

ABSTRACT

The present invention relates to apparatus and methods for quantitative protein design and optimization.